National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

UNIVERSITY OF ALBERTA

MODELING CHILDREN'S PERFORMANCE ON ARITHMETIC WORD PROBLEMS WITH THE LINEAR LOGISTIC TEST MODEL

BY

DOGONI CISSE ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

Edmonton, Alberta

FALL 1994

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-95165-6

Canada

Name DOGONI CISSÉ

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

PSYCHOLOGY
SUBJECT TERM

0 5 2 5
SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture	0729
Art History	0377
Cinema	0900
Dance	0378
Fine Arts	0357
Information Science	0723
Journalism	0391
Library Science	0399
Mass Communications	0708
Music	0413
Speech Communication	0459
Theater	0465

EDUCATION

General	0515
Administration	0514
Adult and Continuing	0516
Agricultural	0517
Art	0273
Bilingual and Multicultural	0282
Business	0688
Community College	0275
Curriculum and Instruction	0727
Early Childhood	0518
Elementary	0524
Finance	0277
Guidance and Counseling	0519
Health	0680
Higher	0745
History of	0520
Home Economics	0278
Industrial	0521
Language and Literature	0279
Mathematics	0280
Music	0522
Philosophy of	0998
Physical	0523
Psychology	[✓ 0525]
Reading	0535
Religious	0527
Sciences	0714
Secondary	0533
Social Sciences	0534
Sociology of	0340
Special	0529
Teacher Training	0530
Technology	0710
Tests and Measurements	0288
Vocational	0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language	
General	0679
Ancient	0289
Linguistics	0290
Modern	0291
Literature	
General	0401
Classical	0294
Comparative	0295
Medieval	0297
Modern	0298
African	0316
American	0591
Asian	0305
Canadian (English)	0352
Canadian (French)	0355
English	0593
Germanic	0311
Latin American	0312
Middle Eastern	0315
Romance	0313
Slavic and East European	0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy	0422
Religion	
General	0318
Biblical Studies	0321
Clergy	0319
History of	0320
Philosophy of	0322
Theology	0469

SOCIAL SCIENCES

American Studies	0323
Anthropology	
Archaeology	0324
Cultural	0326
Physical	0327
Business Administration	
General	0310
Accounting	0272
Banking	0770
Management	0454
Marketing	0338
Canadian Studies	0385
Economics	
General	0501
Agricultural	0503
Commerce-Business	0505
Finance	0508
History	0509
Labor	0510
Theory	0511
Folklore	0358
Geography	0366
Gerontology	0351
History	
General	0578
Ancient	0579
Medieval	0581
Modern	0582
Black	0328
African	0331
Asia, Australia and Oceania	0332
Canadian	0334
European	0335
Latin American	0336
Middle Eastern	0333
United States	0337
History of Science	0585
Law	0398
Political Science	
General	0615
International Law and Relations	0616
Public Administration	0617
Recreation	0814
Social Work	0452
Sociology	
General	0626
Criminology and Penology	0627
Demography	0938
Ethnic and Racial Studies	0631
Individual and Family Studies	0628
Industrial and Labor Relations	0629
Public and Social Welfare	0630
Social Structure and Development	0700
Theory and Methods	0344
Transportation	0709
Urban and Regional Planning	0999
Women's Studies	0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture	
General	0473
Agronomy	0285
Animal Culture and Nutrition	0475
Animal Pathology	0476
Food Science and Technology	0359
Forestry and Wildlife	0478
Plant Culture	0479
Plant Pathology	0480
Plant Physiology	0817
Range Management	0777
Wood Technology	0746
Biology	
General	0306
Anatomy	0287
Biostatistics	0308
Botany	0309
Cell	0379
Ecology	0329
Entomology	0353
Genetics	0369
Limnology	0793
Microbiology	0410
Molecular	0307
Neuroscience	0317
Oceanography	0416
Physiology	0433
Radiation	0821
Veterinary Science	0778
Zoology	0472
Biophysics	
General	0786
Medical	0760

EARTH SCIENCES

Biogeochemistry	0425
Geochemistry	0996
Geodesy	0370
Geology	0372
Geophysics	0373
Hydrology	0388
Mineralogy	0411
Paleobotany	0345
Paleoecology	0426
Paleontology	0418
Paleozoology	0985
Palynology	0427
Physical Geography	0368
Physical Oceanography	0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences	0768
Health Sciences	
General	0566
Audiology	0300
Chemotherapy	0992
Dentistry	0567
Education	0350
Hospital Management	0769
Human Development	0758
Immunology	0982
Medicine and Surgery	0564
Mental Health	0347
Nursing	0569
Nutrition	0570
Obstetrics and Gynecology	0380
Occupational Health and Therapy	0354
Ophthalmology	0381
Pathology	0571
Pharmacology	0419
Pharmacy	0572
Physical Therapy	0382
Public Health	0573
Radiology	0574
Recreation	0575
Speech Pathology	0460
Toxicology	0383
Home Economics	0386

PHYSICAL SCIENCES

Pure Sciences

Chemistry	
General	0485
Agricultural	0749
Analytical	0486
Biochemistry	0487
Inorganic	0488
Nuclear	0738
Organic	0490
Pharmaceutical	0491
Physical	0494
Polymer	0495
Radiation	0754
Mathematics	0405
Physics	
General	0605
Acoustics	0986
Astronomy and Astrophysics	0606
Atmospheric Science	0608
Atomic	0748
Electronics and Electricity	0607
Elementary Particles and High Energy	0798
Fluid and Plasma	0759
Molecular	0609
Nuclear	0610
Optics	0752
Radiation	0756
Solid State	0611
Statistics	0463

Applied Sciences

Applied Mechanics	0346
Computer Science	0984
Engineering	
General	0537
Aerospace	0538
Agricultural	0539
Automotive	0540
Biomedical	0541
Chemical	0542
Civil	0543
Electronics and Electrical	0544
Heat and Thermodynamics	0348
Hydraulic	0545
Industrial	0546
Marine	0547
Materials Science	0794
Mechanical	0548
Metallurgy	0743
Mining	0551
Nuclear	0552
Packaging	0549
Petroleum	0765
Sanitary and Municipal System Science	0554
System Science	0790
Geotechnology	0428
Operations Research	0796
Plastics Technology	0795
Textile Technology	0994

PSYCHOLOGY

General	0621
Behavioral	0384
Clinical	0622
Developmental	0620
Experimental	0623
Industrial	0624
Personality	0625
Physiological	0989
Psychobiology	0349
Psychometrics	0632
Social	0451

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: DOGONI CISSE

TITLE OF THESIS: MODELING CHILDREN'S PERFORMANCE ON ARITHMETIC WORD PROBLEMS WITH THE LINEAR LOGISTIC TEST MODEL

DEGREE: DOCTOR OF PHILOSOPHY

YEAR THIS DEGREE GRANTED: FALL 1994

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form without the author's prior written permission.

P. O. Box 5055
Bamako (Republic of Mali)
West Africa

October 5, 1994

"It is only a slight exaggeration to describe the test theory that dominates educational measurement today as the application of 20th century statistics to 19th century psychology... The application of modern statistical methods with modern psychological models constitutes the foundation of a new test theory." (Mislevy, 1993; p. 19)

"Applications of mathematics are always complicated by the obligation to be true to the subject matter treated as well as to the mathematics." (Mosteller & Tukey, 1977; p. 1)

"... if we are serious in our attempts to measure, we must examine every application to see how well each set of responses corresponds to our model expectations." (Wright & Stone, 1979; p. 66)

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled MODELING CHILDREN'S PERFORMANCE ON ARITHMETIC WORD PROBLEMS WITH THE LINEAR LOGISTIC TEST MODEL submitted by DOGONI CISSE in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

Maguire
Dr. Thomas O. Maguire

Dr. Donald C. Heth

S. Hunka
Dr. Stephen M. Hunka

Thomas Kieren
Dr. Thomas E. Kieren

RH Short
Dr. Robert H. Short

Dr. Clement Dassa

June 27, 1994

This thesis is dedicated to
the memory of my brother,
Sounkalo Amadou Cisse
(1944 - 1991)

to my parents
(for enduring my absence for so long)

to my wife and children
(for always loving and believing in me)

Abstract

This study employed a psychometric modeling approach to investigate the importance of six problem features in determining the difficulty of arithmeic word problems. Three of the problem features related to logico-mathematical knowledge and the other three to linguistic knowledge.

Forty children from each of Grades 1, 2, and 3 were asked to solve 32 addition and subtraction word problems consisting of eight problems from each of four problem categories (change, combine, compare, and equalize).

The linear logistic test model (LLTM) was used to explore the influence of the six problem features (or complexity factors) on children's performance. The LLTM is a constrained Rasch model that combines a mathematical model of component processes or task features and a psychometric model of individual differences at the item level.

As an item-response theory (IRT) model, the LLTM makes strong assumptions about test data. No evidence was found that the assumptions of this model were violated by the test data, which allowed the application of the model to the data.

The six complexity factors were found to be related to problem difficulty, as defined by the Rasch item difficulties, for the whole sample and for each grade level. Each of the two different sets of complexity factors (i. e., the logico-mathematical and the linguistic) was also related to problem difficulty, but not as strongly as the full set of complexity factors. In other words, the full cognitive model, consisting of all six complexity factors had significantly more predictive power than each of the two submodels consisting of the two different sets of complexity factors.

Only three complexity factors contributed to problem difficulty, namely, knowledge of part-whole relationships, consistency of language, and double-role vs. single-role counters.

The implications of these findings for educational practice and for research in children's performance on arithmetic word problems were discussed. It was also concluded that the LLTM, as illustrated in the present research, could be a useful modeling tool in many research areas.

Acknowledgements

I am indebted to the People of Mali for the fellowhip award which allowed me to pursue my graduate studies in North America.

I also owe a debt of gratitude to the Department of Educational Psychology, the Department of Psychology, the Division of Educational Research Services (DERS), and the Center for Research in Applied Measurement and Evalution (CRAME), all at the University of Alberta, for providing me substantial support for many years in the form of graduate research and teaching assistanships. I thank the Faculty of Graduate Studies and Research at the University of Alberta for the dissertation fellowship, the Andrew Stewart Graduate Prize, and the travel grant I received while I was a doctoral student at the university.

I would like express my appreciation to my research supervisor, Dr. T. O. Maguire, for his guidance, help, and counsel during the preparation of this thesis. I also thank Drs. C. D. Heth, S. M. Hunka, T. E. Kieren, and R. H. Short, for their willingness to serve on my supervisory and/or examination committees. I am thankful to Dr. Clement Dassa of the University of Montreal, who served as the external reader of the dissertation. The comments and suggestions made by all these scholars contributed greatly to the final product of my doctoral research.

I wish to express my sincere thanks to the Edmonton Public School Board, to the children of Baturyn Elementary School who participated in the present research, to the principal and teachers of the school for their support of the project, and to the two individuals who provided assistance with data collection. Their willingness to help out

was instrumental in the successful completion of the research reported in my dissertation.

I would like to acknowledge Dr. C. J. Brainerd and Dr. J. H. Bisanz for their guidance and advice during the early years of my doctoral studies.

Finally, I would like to express my thanks to my wife and children, my parents and relatives, and all my friends, for their sustained support and encouragement, which contributed so much to my sanity during so many years.

Table of Contents

Chapter Page

I. INTRODUCTION 1

  A. Background of the problem 1

  B. The problem 4

II. LITERATURE REVIEW 11

  A. Research trends in children's arithmetic 11

    1. Beginnings of research into children's arithmetic: The "ancients" revisited 11

    2. Research into children's arithmetic: the interim years 14

    3. Research into children's arithmetic: the current scene 17

  B. Models of children's performance on arithmetic word problems 23

    1. Overview of the models 23

    2. A schema-based model 24

    3. A model emphasizing problem-solving procedures . . 30

  C. Role of knowledge in children's solutions to arithmetic word problems 35

    1. Knowledge and cognitive performance 35

    2. Knowledge and children's mathematical cognition . . 38

D. Latent trait theory and a model for component processes . . 45

1. Latent trait theory 45

2. The linear logistic test model 64

E. Summary and rationale for the present study 70

III. METHOD 72

A. Subjects 72

B. Tasks and materials 72

C. Procedure 79

D. Data-analysis strategies and presentation of results 82

IV. RESULTS 91

A. Preliminary analysis: sizing up the measuring instrument . . 91

B. Main analysis 98

1. Describing children's performance on the problems . 98

2. Assessing goodness-of-fit of the model to the data . 102

3. Applying the LLTM to the data 110

V. DISCUSSION 127

REFERENCES 134

List of Tables

Table — Page

1. Classification of addition and subtraction word problems according to the RGH model 26

2. Test problems used in the study (Frame 1) 76

3. Test problems used in the study (Frame 2) 78

4. Item sequence, item code, and the Q matrix 90

5. Classical item analysis results for change problems 93

6. Classical item analysis results for combine problems 94

7. Classical item analysis results for compare problems 95

8. Classical item analysis results for equalize problems 96

9. Descriptive information about classical psychometric indexes of test items 97

10. Subtest statistics 97

11. Subtest correlation matrix 98

12. Proportions of correct solutions as a function of grade 99

13. First four eigenvalues and percentages of variance accounted for (principal-components analysis) 103

14. First four eigenvalues and percentages of variance accounted for (principal-axes factoring) 103

15. First four eigenvalues for observed and random data from parallel analysis 104

16. Description of residual matrices after fitting linear and non-linear factor analysis models 105

17. Rasch item difficulties and their standard errors as estimated by BICAL 109

18. η parameter estimates and their standard errors 111

19. Normalized LLTM and Rasch model estimates 113

20. Log-likelihoods (full model) 114

21. Full model η estimates, standard errors, and t values (whole sample) 119

22. Full model η estimates, standard errors, and t values (Grade 1) 119

23. Full model η estimates, standard errors, and t values (Grade 2) 119

24. Full model η estimates, standard errors, and t values (Grade 3) 120

25. Submodel 1 η estimates, standard errors, and t values (whole sample) 120

26. Submodel 1 η estimates, standard errors, and t values (Grade 1) 120

27. Submodel 1 η estimates, standard errors, and t values (Grade 2) 120

28. Submodel 1 η estimates, standard errors, and t values (Grade 3) 121

29. Submodel 2 η estimates, standard errors, and t values (Whole sample) 121

30. Submodel 2 η estimates, standard errors, and t values (Grade 1) 122

31. Submodel 2 η estimates, standard errors, and t values (Grade 2) 122

32. Submodel 2 η estimates, standard errors, and t values (Grade 3) 122

33. Regression analysis results for Grade 1 124

34. Regression analysis results for Grade 2 124

35. Regression analysis results for Grade 3 124

List of Figures

Figure Page

1. 2-PL ICCs for four items 54

2. 1-PL ICCs for four items 54

3. 3-PL ICCs for six items 54

4. 3-PL ICCs for three items (same a and c values, different β values) 58

5. 3-PL ICCs for three items (same β and c values, different a values) 58

6. 3-PL ICCs for three items (same a and β values, different c values) 58

I. INTRODUCTION

A. Background of the problem

Mary has 6 marbles. John has 2 marbles. How many marbles does John have less than Mary?

This problem exemplifies the many types of problems that researchers have used to investigate children's performance on one-step arithmetic word problems. As reported by Cummins, Kintsch, Reusser, and Weimer (1988), only 29% of the first grade children in their study could solve this problem correctly. However, when the same problem was stripped of the features making it a word problem, that is, when it was presented in a numeric format as a number sentence (i.e., 6 - 2 = ?), all the children in their study could solve it correctly.

This is by no means an isolated finding. In effect, researchers and educators have long recognized that one of the most troublesome areas of mathematical instruction at the elementary school level involves arithmetic word-problem solving. At present, it is known that, in the United States of America, children perform 10 to 30 percent worse on arithmetic word problems than on comparable problems presented as number sentences. The National Assessment of Educational Progress (NAEP) has drawn the attention of educators, policy makers, and the general public to children's poor performance on arithmetic word problems. For instance, the second NAEP mathematics assessment at the elementary school level revealed that "only 28%, as compared to 46% on the previous assessment, could solve a word problem involving the product of a 2-digit number by a 1-digit number" (Carpenter, Kepner, Corbitt, Linquist, & Reys, 1980;

p. 12). As explicitly stated in the previous quote, there was even a downward trend in children's ability to solve arithmetic word problems, during the period covered by the second NAEP. Indeed, children's difficulties in arithmetic word problem solving have been referred to as a "national crisis" (Kameenui & Griffin, 1989; p. 575).

It is not surprising then that the topic of arithmetic word-problem solving during the elementary school years has been a major concern of both researchers and educators for many years. Furthermore, research in this area has been stimulated by the idea that "problem solving must be the focus of school mathematics in the 1980s" (National Council of Teachers of Mathematics, 1980; p. 2).

The main objective of much research dealing with children's performance on arithmetic word problems has been to determine the sources of children's difficulties with these problems. This research thrust has led many investigators to focus on the knowledge structures and cognitive processes involved in children's performance on arithmetic word problems. To date, several process models of children's arithmetic word-problem solving have been proposed (e.g., Briars & Larkin, 1984; Kintsch & Greeno, 1985; Riley & Greeno, 1988; Riley, Greeno, & Heller, 1983). The knowledge required for competent performance on arithmetic word problem is often implicit. The intent of the research on knowledge structures is to make that knowledge explicit by specifying it in detail, sometimes in the form of computer-simulation models.

In their attempt to build models of the knowledge underlying various levels of problem-solving ability in arithmetic word problems, researchers have taken two different approaches. The first approach involves making detailed analyses of children's solutions to arithmetic word problems. The second approach involves making detailed analyses of the mathematical content of the problems to be solved. Although the two approaches can be distinguished, they are not mutually exclusive, as they have been used together by several researchers.

The research on knowledge structures has resulted in rich descriptions of how children solve arithmetic word problems. These descriptions include not only the various types of errors or incorrect applications children make, but also the kinds of correct or appropriate procedures children employ and sometimes invent to solve various arithmetic word problems. The research has resulted in a typology of one-step addition and subtraction word problems, and some consensus has been reached among researchers about its adequacy in characterizing the problems.

Investigators studying children's performance on arithmetic word problems have concluded that skilled performance depends heavily on the kinds of knowledge structures, or schemata, that children bring to the problems. This schema notion is a "close relative" of the Piagetian idea that people's cognitive structures determine the nature and power of their problem-solving abilities. Clearly, in both accounts, the available structures, or schemata, limit the range of problems a person can solve. Furthermore, the aspects of a problem that are attended to, the interpretations that they receive, and the ways in which they are approached, are believed to be heavily dependent on the interpretive structures, or schemata, brought to the problem.

B. The problem

The general view that the knowledge structures available to children determine the course and the outcome of their arithmetic word-problem solving is fairly uncontroversial. There is ample evidence, for example, that expert and novice problem solvers differ in the schemata they possess, and hence reason differently, and adopt different strategies when solving problems. Although the research evidence (e.g., Chi, Feltovich, & Glaser, 1981; Larkin, McDermott, Simon, & Simon, 1980) has come mostly from other problem-solving domains (e.g., chess, physics), it seems intuitively compelling that the same should hold true in the domain of arithmetic word-problem solving. Indeed, it is widely accepted among researchers that arithmetic word problems will prove difficult to children when the capacities required to process them are not yet in the children's repertoires. The change of problem difficulty patterns with age has led many investigators to adopt a Piagetian view, that is, a *competence* interpretation of solution characteristics.

However, the question concerning which capacities develop over time to improve children's solution performance is still open to debate among investigators. A quick perusal of the relevant literature quickly reveals that two main groups of researchers hold contrasting views concerning the question. For one group of investigators, improvement in performance on arithmetic word problems can be attributed to the development of logico-mathematical knowledge. In other words, children become capable of solving certain problems because they have acquired the conceptual knowledge required to solve them. Briars and Larkin (1984), Carpenter and Moser (1982, 1983), Nesher, Greeno,

& Riley (1982), Riley and Greeno (1988), Riley et al. (1983), and Vergnaud (1982) are among the proponents of the logico-mathematical view.

According to Riley and her colleagues, poor performance on certain word problems reflects a lack of sufficient knowledge concerning part-whole set relations. For these investigators, the development of the part-whole schema in children is the crucial determinant of solution success on the most difficult addition and subtraction word problems. On these problems, text phrases do not provide key words that problem solvers can map onto solution procedures, as in the case of using a counting procedure when the word "altogether" is heard. Hence, Riley and her colleagues view the part-whole schema as the hallmark of the most advanced level of arithmetic word-problem solving. They argue that the part-whole schema enables children to deal with problems more flexibly by interpreting the semantic structures of different arithmetic word problems in terms of parts and wholes. In other words, the acquisition of the part-whole schema has an indirect effect on children's problem-solving skills in that it improves children's problem representations.

Other proponents of the logico-mathematical view (e.g., Briars & Larkin, 1984; Nesher et al., 1982) contend that the acquisition of appropriate conceptual knowledge exerts a more direct effect on problem-solving strategies. For Briars and Larkin, the acquisition of logical knowledge (e.g., single-role and double-role counters, subset equivalence, and time reversal) directly determines the strategies chosen by children in solving arithmetic word problems. However, it is important to realize that Briars and Larkin go beyond a pure logico-mathematical view, as they also think that, for some problems involving comparisons, children's knowledge of "consistent-comparison" and "conflict-comparison" language is critical to problem-solving success. For Nesher et al. (1982), the development of knowledge concerning sets, logical operations, and mathematical operations are aspects of conceptual knowledge that are critical in the development of arithmetic word-problem solving skills.

There are other investigators who espouse a linguistic development view. They argue that certain word problems are difficult to solve "because they employ linguistic forms that do not readily map onto children's conceptual knowledge structures" (Cummins et al., 1988; p. 407). They contend that a child may, for example, understand part-whole set relations and yet be uncertain as to how the comparative verbal form (e.g., more than) in some arithmetic word problems maps onto them. In that case, children's problems with word problems are really with the understanding of such verbal forms. This is clearly a *performance* interpretation. Investigators that have reported evidence consistent with that view include Cummins (1991), Cummins et al. (1988), De Corte, Verschaffel, and DeWin (1985), Davis-Dorsey, Ross, and Morrison (1991), Hudson (1983), Lewis (1989), Lewis and Mayer (1987), and Stern (1993).

It should be observed that the logico-mathematical view and the linguistic development view are not completely antithetical positions. Rather, they can be viewed as complementary, since investigators favoring the linguistic development view do not claim that part-whole knowledge is not critical to successful problem-solving performance. Proponents of the linguistic development view only emphasize that, for some problems, "solution failures result from missing or inadequate mappings of verbal expressions to part-whole structures leading to a failure to access part-whole knowledge" (Cummins, 1991; p. 265). Thus, these researchers explain the facilitative effect of rewording reported by several investigators (e.g., De Corte et al., 1985; Hudson, 1983) by suggesting an increase in the probability of successfully accessing part-whole knowledge.

The linguistic development view, then, proposes that word problems containing certain verbal forms constitute tests of verbal sophistication (i.e., linguistic development) as well as of conceptual maturity (i.e., logico-mathematical knowledge). Consequently, when errors occur on these problems, one is confronted with the problem of determining their source; specifically, are errors the result of deficiencies in language-comprehension skills, part-whole knowledge, or both? Proponents of the linguistic developmental view contend that, most often, what is implicated in children's failures is not lack of conceptual knowledge, but deficiencies in linguistic knowledge.

In summary, some proponents of the logico-mathematical view of the development of problem-solving skills in arithmetic word problems consider part-whole knowledge as a necessary and sufficient component of problem-solving success on certain arithmetic word problems, when text phrases provide no cues to appropriate solution strategies. Other proponents of the same view emphasize the crucial role of the understanding of other concepts; for example, subset equivalence and time reversal. On the opposite side, proponents of the linguistic development view contend that, although part-whole knowledge is necessary for competent performance on some arithmetic word problems, it may not be sufficient in cases when performance factors (e.g., inability to understand the language) hinder access to that knowledge, making its use impossible.

There are problems with both views of the development of competence in solving arithmetic word problems. The difficulty with the logico-mathematical view is that, at present, empirical investigations about its propositions are either lacking or have reported ostensibly conflicting findings. First, concerning the lack of evidence, the present writer is not aware of any studies that have investigated the hypothesized role of children's understanding of the concept of re-representation, for example, in performance on arithmetic word problems. Second, with regard to discrepant empirical results, some have been reported about the role of part-whole knowledge in performance on arithmetic word problems (e.g., Dean & Malik, 1986; Fuson & Willis, 1989; Morales, Shute, & Pellegrino, 1985; Rathmell, 1986; Tamburino, 1982; Willis & Fuson, 1988; Wolters, 1983). The problem with the linguistic development view is that it raises to the status of an established fact the *hypothesis* that part-whole knowledge plays a crucial role in children's performance on some arithmetic word problems, although other factors also may play a role.

Current cognitive theories of the development of problem-solving skill in arithmetic word problems refer to knowledge of part-whole set relations as the part-whole schema. The part-whole schema is an abstract knowledge structure that specifies that any quantity (the whole) can be partitioned (into the parts), as long as the combined parts neither exceed nor fall short of the whole. By implication then, the parts make up or are included in the whole. As is the case for every schema, the part-whole schema consists of elements or slots (i.e., the whole and the parts) and relations between them. In the domain of arithmetic word problems the relations are quantitative, i.e.,the part-whole relationships that are of interest are numerical in nature.

The purpose of the present study was to explore the role of conceptual and linguistic knowledge in the development of problem-solving competence in addition and subtraction word problems using a psychometric modeling approach. More specifically, an extension of the one-parameter logistic (1-PL) or Rasch model of item-response theory (IRT), namely, the linear logistic test model (LLTM) was employed to investigate the role of children's conceptual knowledge (as embodied primarily in their understanding of numerical part-whole set relations) and the role of their linguistic understanding of problem statements. The LLTM is one of several modifications of one of the original IRT models. The development of the LLTM was motivated by the need, in problem-solving research, to identify the cognitive components underlying performance on a specific task or to test the adequacy of the response processes or cognitive components hypothesized by a given problem-solving model.

The objectives of the present study were both substantive and methodological. On the substantive side, the purpose of the study was to evaluate the adequacy of models that posit conceptual knowledge and linguistic knowledge as crucial determinants in the development of children's competence in solving arithmetic word problems. On the methodological side, the objective of the study was to illustrate the use of a latent trait model as a technique for investigating the structural features and processing requirements of one-step addition and subtraction word problems.

The next chapter provides a review of related literature as it pertains to (i) research on children's arithmetic, with a focus on their performance on one-step addition and subtraction word problems and (ii) models of modern test theory useful in the investigation of problem-solving processes.

II. LITERATURE REVIEW

The review is organized under four sections. Section 1 presents a historical overview of research dealing with children's arithmetic. Section 2 presents models of children's performance on arithmetic word problems. It also provides a description of the various kinds of arithmetic word problems that have been studied by researchers. Section 3 reviews studies that have investigated the role of knowledge in children's performance on arithmetic word problems, with an emphasis on logico-mathematical knowledge and linguistic knowledge. Section 4 is a methodology review of item-response theory (IRT) models tools for analyzing test data. This section also describes the linear logistic test model (LLTM) and explicates its use as a technique for identifying the components or complexity factors underlying performance in specific task domains.

A. Research trends in children's arithmetic

1. Beginnings of research into children's arithmetic: the "ancients" revisited

Psychological and educational investigations into children's problem-solving processes in arithmetic and their learning of arithmetic have a long history. Studies of children's solutions to arithmetic problems date back to the beginning of the century (e.g., Arnett, 1905; Browne, 1906). These early reports were instrumental in setting the stage for a frenzied research activity dealing with children's arithmetic, as evidenced by the substantial number of writings on the topic during the first half of the century. For example, Brownell (1928, 1947), Buswell and Judd (1925), Clapp (1924), Knight and

Behrens (1928), and Wheeler (1939) were just a few of all those cited in *The Psychology of Mathematics for Instruction* (Resnick & Ford, 1981).

Several reasons can be identified that account for these researchers' interest in children's arithmetic. First, as noted by Carpenter and Moser (1983), because arithmetic is a clearly specifiable domain, researchers strongly believed it could be used to study general principles of learning and instruction. For instance, when Thorndike (1922), in *The Psychology of Arithmetic*, formulated his seven rules for the teaching of arithmetic, the rules were also intended to be his rules for good teaching in general; arithmetic was only taken for convenience. Second, researchers also thought that arithmetic was well suited for the task of providing "a window on basic cognitive processing within a child..." (Carpenter & Moser, 1983; p. 7). Finally, when researchers endeavored to apply psychology to education, very often, arithmetic was the target domain. For example, Thorndike (1922) chose arithmetic in his attempt to apply psychological principles such as associations between stimuli and responses, bonds, and the law of effect, to school learning.

A sizable part of early research into children's arithmetic focused on addition and subtraction and was concerned essentially with the relative difficulty of combinations of small numbers (i.e., from 0 to 9) in these two operations. The main task was to identify addition and subtraction problems children could and could not solve. The ultimate product was a ranking, with respect to relative difficulty, of the basic number facts. This, of course, was not an idle exercise, as it was meant to provide valuable information for curriculum sequencing and instructional time allocation. There was also an interest in some studies (e.g., Knight & Behrens, 1928) in uncovering some of the factors underlying the difficulty of number combinations in addition and subtraction.

For both lines of research the literature was replete with contradictory findings. For example, with respect to the first line of research, different rankings of the relative difficulty of number facts were reported by different studies. Indeed, only one consistent pattern emerged from each line of research. First, almost all studies found that difficulty increases as the numbers in the problem get bigger. Second, there was a clear agreement among various studies that combinations in which the numbers were the same (e.g., 5 + 5 = ?) were easier than other combinations of numbers of comparable size (e.g., 4 + 6 = ?).

It should be noted that early studies of children's arithmetic were not limited to the relative difficulty of number facts. There were also investigations concerned with arithmetic story problems, or word problems, such as those of Brownell and Stretch (1931), Hydle and Clapp (1927), Kramer (1933), all cited in Resnick and Ford (1981). Here also much of the research effort was directed towards ranking problems on the basis of various measures of problem complexity. Researchers uncovered many features of the problems, most of them related to context and wording, that had marked effects on performance, for instance, the level of vocabulary used, the number of unfamiliar and nonessential elements, and the intrinsically interesting nature of the story problems. There was, however, no overall agreement among researchers on the relevance of all the identified factors. A case in point was the exchange between Hydle and Clapp (1927) and Brownell and Stretch (1931) on the critical role of the familiarity of problem situations.

In summary, early research on children's arithmetic was mainly concerned with problem difficulty and had, as objectives, the ranking of problems with respect to relative difficulty and the identification of factors underlying problem difficulty. The problems studied were number combinations in addition and subtraction, as well as story problems. As will be seen later, these topics still interest researchers in the field.

2. Research into children's arithmetic: the interim years

The research on the relative difficulty of number combinations in addition and subtraction continued unabated throughout the 1950s and 1960s and well into the 1970s. There was, however, a major difference with previous research concerned mainly with rankings based on the relative difficulty of problems. In effect, it became progressively clear to investigators that difficulty rankings were influenced, to a large extent, by the conditions under which they were obtained and by the techniques by which they were determined. One direct implication of this fact was that the notion of the intrinsic difficulty of number facts that had guided previous research was now viewed as misleading, or at least counterproductive. Consequently, although researchers continued to be interested in the relative difficulty of different number combinations in addition and subtraction, they shifted their emphasis to the structural features of the problems and the conditions under which the problems were administered. To illustrate, one of the structural variables considered by investigators was the type of open sentences used in the problems, which is determined by the identity of the unknown. By changing the identity of the unknown, researchers have generated six open sentences problems for both addition and subtraction. The type of open sentence has proved to be a "fruitful" variable, in that various results were found consistently across studies (e.g., Grouws, 1972; Weaver, 1971) investigating the effects of the type of open sentence on the difficulty of number facts. For instance, studies have repeatedly shown that canonical addition sentences ($a + b = ?$) are easier than noncanonical addition sentences ($a + ? = 6$, $? + b = 6$). They have also shown that the missing minuend sentence ($? - b = c$) is more difficult than are the other five subtraction sentences. (For the other results about the effects of the type of open sentence, see Carpenter and Moser, 1983.)

The new emphasis on the structural features and the conditions of administration of number facts was extended to research on word problems as well. Many studies (e.g., Hebbeler, 1977; Steffe, 1970; Steffe & Johnson, 1971) investigated the effects of concrete and pictorial aids on children's performance on addition and subtraction word problems. These concrete and pictorial aids were found to have facilitative effects on performance. Other studies (e.g., Jerman, 1973; Jerman & Mirman, 1974; Jerman & Rees, 1972; Suppes, Loftus, & Jerman, 1969) attempted to build regression models to predict the difficulty of different problems. In these models, various structural and syntactic variables such as problem length and sentence complexity, were used as predictor variables. Note, however, that this endeavor was clearly atheoretical because the models constructed by researchers were not based on any explicitly stated theories of arithmetic performance.

Alongside the foregoing studies, there were investigations concerned with the relationship between the development of Piagetian logical reasoning abilities in children and children's performance in arithmetic. The impetus for this research was provided by *The Child's Conception of Number* (Piaget, 1952). Piaget proposed that addition and subtraction require inversion reversibility, which is not achieved, according to him, until the period of concrete operations. This contention raised the issue of the usefulness of Piagetian tasks as readiness measures, that is, whether they were good empirical predictors of children's ability to learn arithmetic and other mathematical concepts and skills. Hiebert and Carpenter (1982) provided a good review of that research. As observed by these authors, some investigators (e.g., Freyberg, 1966; Kaufman & Kaufman, 1972) adopted a global approach whereby a battery of Piagetian tasks and a test of mathematics achievement were administered to primary grade children and correlations between scores were computed. Other researchers (e.g., Dimitrovsky & Almy, 1975; Dodwell, 1961) considered the relationship between specific measures of logical reasoning (e.g., number conservation) and children's skill in particular mathematical domains (e.g., arithmetic computation). Here, the task was not to calculate a single correlation coefficient; rather, the goal was to compare the performance of children at different developmental levels. For instance, a comparison of addition and subtraction computations was done among conservers, nonconservers, and those at a transitional level in conservation acquisition.

In spite of the uniformly high positive correlations between performance on Piagetian tasks and various measures of arithmetic proficiency, Piagetian tasks appeared to have limited use as readiness measures. The key result was that children who did not pass tests of conservation, seriation, transitivity, or class inclusion, could nonetheless learn to add and subtract.

To sum up, the trends set during the previous decades of research into children's arithmetic were nurtured during the three decades from the 1950s to the 1970s. Specifically, investigators continued to show a marked interest in the relative difficulty of number combinations in addition and subtraction. However, in their attempt to account for problem difficulty, they were now emphasizing the structural features of the problems and their conditions of administration. Other important investigations during these decades were stimulated by Piaget's ideas on the development of number concepts in children. Piaget (1952) underestimated the significance of basic quantitative skills such as counting in the acquisition of skill in arithmetic. Rather, he posited prerequisite logical abilities to children's understanding of arithmetic. Researchers who have followed Piaget's lead also have tended to de-emphasize the processes and strategies used by children to solve arithmetic problems. Instead, they have focused on several Piagetian logical abilities, which they thought, intuitively at least, might be necessary for learning arithmetic. The empirical evidence was not supportive of these claims.

3. Research into children's arithmetic: the current scene

A large part of the more recent and the current research into children's arithmetic differs from the older studies, both in its techniques and its theoretical approach. As observed by Romberg (1982), "to a large extent, the many studies on addition and subtraction represent an eclectic morass" (p. 1). This was clearly an accurate description of older studies on children's arithmetic because these studies were not guided by any unified theoretical perspective.

However, currently there exists a research consensus in children's arithmetic, a view echoed by Ashcraft (1982) and Romberg (1982). Both authors observed that a paradigm, in Kuhn's (1970) sense of the term, has emerged in psychological research in general and in research into children's arithmetic in particular. As observed by Ashcraft (1982), "what has emerged is a psychology of cognition, a paradigm which relies heavily on time-based, that is to say, 'chronometric', methods for its models and theories..." (p. 214). What Ashcraft is referring to is clearly the information-processing perspective, which is arguably today the leading general framework for the study of cognition and cognitive development. According to Siegler (1983), the appeal of the information-processing paradigm derives from many sources, including its general perspective on human beings, the usefulness of its language for characterizing cognition, its arsenal of powerful methodologies and techniques for studying cognition, and the issues it addresses. (For a detailed treatment of the information-processing perspective, see Kail & Bisanz, 1982; Klahr, 1989; Lachman, Lachman, & Butterfield, 1979; and Siegler, 1983.)

Certainly, specific results from information-processing studies of arithmetic are important; however, the most potentially impactful ideas are the core constructs underlying the approach, such as *representation*, *processes*, *limited attentional resources*, and *automaticity*. One distinguishing feature is that humans are characterized as limited capacity manipulators of symbols. Another feature concerns the methods developed to test information-processing theories, which are detailed models of performance and development of skill in specific task domains (e.g., children's arithmetic). The models are often computer-simulation models (e.g., Briars & Larkin, 1984; Riley & Greeno, 1988), a fact which is consistent with the view that the computer can be used productively as a metaphor to study human cognition. They are usually evaluated using various kinds of data and types of analysis. The latter include (1) *chronometric methods*, which utilize patterns of subjects' reaction times, as these are assumed to reflect the time course of information processing; (2) *error analyses*, which emphasize patterns of correct answers and errors to clarify subjects' conceptual understanding; and (3) *protocol analyses,* which rely on subjects' verbalizations to uncover the strategies used to solve cognitive tasks.

A process approach: Most current research into children's arithmetic shows an obvious concern for internal cognitive processes. The goal for research is to identify not only which problems children can solve, but also the processes and skills they use to solve them. The focus is not on underlying mental abilities such as class inclusion or number conservation, but on the processes children use to solve arithmetic problems. Indeed, Romberg (1982) thought that this is the essence of the new paradigm when he wrote: "The emerging general paradigm is to formulate precise models of the cognitive processes used by subjects when carrying out specific tasks and how those processes change over time" (p. 3).

It should also be observed that current research provides a much clearer specification of both problem structure and children's solution strategies than was found in earlier studies (e.g., Brownell, 1928; Gibb, 1956). Furthermore, as pointed out by Carpenter and Moser (1983), in current research, investigators attempt to describe both problem structure and children's strategies in such a way that a clear connection can be drawn between different problems and the strategies used to solve them. Finally, researchers are no longer content to simply describe the overt strategies that children use; the goal is now to describe the cognitive processes related to the overt strategies.

A problem-solving focus: The learning of school mathematics has now been widely recognized as an opportunity for students to learn about problem solving. Pleas have been made that time and effort be redirected from drill and practice on computations to the development of problem-solving strategies. The focus of school mathematics on problem solving accounts for the increasing interest of many investigators in information-processing approaches to problem solving.

Early information-processing research was focused on the study of puzzle-like problems (Newell & Simon, 1972). More recently, however, interest has turned towards problem solving in complex subject-matter domains such as physics (e.g., Larkin, McDermott, Simon, & Simon, 1980; Simon & Simon, 1978), computer programming (e.g., Jeffries, Turner, Polson, & Atwood, 1981), geometry (e. g., Greeno, 1978), and arithmetic word problems (e.g., Briars & Larkin, 1984; Riley & Greeno, 1988; Riley et al., 1983). Very often, the problems examined are fairly standard textbook problems of the kind students typically encounter in school.

Mental arithmetic and arithmetic word problems: Two major topics for research into Children's mathematical cognition: The literature on current research into children's mathematical problem solving is voluminous, and researchers have been interested in various questions. However, even a casual perusal of the literature will quickly reveal that children's arithmetic has continued to attract the attention of investigators. A multitude of studies have been concerned with mental arithmetic, mostly addition and subtraction (e.g., Ashcraft, 1982; Ashcraft & Fierman, 1982; Ashcraft, Fierman, & Bartolotta, 1984; Baroody, 1984; Brainerd, 1983; Groen & Parkman, 1972; and Haman & Ashcraft, 1985, 1986), but also multiplication, albeit to a lesser degree (e.g., Campbell & Graham, 1985). As observed by Campbell and Graham, a concern with the basic arithmetic operations is motivated primarily by one consideration: Despite more than a half century of research on the topic, no firm conclusions can be drawn regarding how basic arithmetic should be taught to children, and why the acquisition of simple number facts presents a real challenge to a large number of children.

Children's performance on arithmetic word problems also has continued to be a major focus of research into children's mathematical cognition. Here again, the practical concern of improving school learning is among the chief motivating factors. Carpenter (1985) was unequivocal about the other reasons for researchers' sustained interest in arithmetic word problems when he wrote:

> *The domain of problems is simple enough that differences between problems can be specified with a reasonable degree of clarity. On the other hand, the domain is rich enough to provide a variety of problems, solution strategies, and errors. By the same token, children's solution processes appear to be simple enough to provide some hope of understanding and modeling them but complex enough to be interesting.* (p. 17)

Studies of children's performance on arithmetic word problems are legion. For instance, taxonomies of arithmetic word problems have been proposed and a common scheme for classifying problems has virtually emerged (e.g., Carpenter, 1985; Carpenter & Moser, 1982, 1983; Nesher, Greeno, & Riley, 1982; Riley & Greeno, 1988; Riley et al., 1983; Vergnaud, 1982). Models of children's solutions to arithmetic word problems have been proposed and refined (Briars & Larkin, 1984; Kintsch & Greeno, 1985; Riley & Greeno, 1988; Riley et al., 1983). Various suggestions have been made regarding children's shortcomings and misconceptions when solving arithmetic word problems (De Corte & Verschaffel, 1985, 1987; Lewis & Mayer, 1987). Teaching experiments have been conducted in an attempt to boost children's performance (Fuson & Willis, 1989; Willis & Fuson, 1988; Wolters, 1983). Children's sorting patterns of arithmetic word problems have been examined to clarify their problem representations and to relate these representations to their problem solutions (Morales, Shute, & Pellegrino, 1985). Their skills in detecting errors in word-problem statements have been investigated (Haneghan, 1990). Performance on arithmetic word problems has been considered from the point of view of developmental sequences (Dean & Malik, 1986; Nesher et al., 1982). Arithmetic word problems found in American and Soviet textbooks have been compared with regard to difficulty (Stigler, Fuson, Ham, & Kim, 1986). Several interpretations of problem difficulty have been investigated, such as those based on working memory (Fayol, Abdi, & Gombert, 1987), problem wording (Davis-Dorsey, Ross, & Morrison, 1991; De Corte, Verschaffel, & Dewin, 1985; Hudson, 1983), and involvement of a specific knowledge schema (Willis & Fuson, 1988; Wolters, 1983). Clearly, as already mentioned, children's performance on arithmetic word problems has provided a ground for a vigorous research activity.

B. Models of children's performance on arithmetic word problems

1. Overview of the models

A view that has been expressed in recent information-processing investigations of the development of problem solving in children is that performance on most tasks and acquisition of skill on those tasks involve an interaction between *conceptual* knowledge and *procedural* knowledge. Conceptual knowledge has been defined as knowledge of relationships, and procedural knowledge as knowledge of rules, principles, and algorithms in a problem domain. This view is now prominent in current analyses of children's counting and numerical competence (Briars & Siegler, 1984; Gelman & Meck, 1983; Gelman, Meck, & Merkin, 1986; Hiebert, 1985) and in models of the development of problem-solving skills in arithmetic word problems (Briars & Larkin, 1984; Riley & Greeno, 1988; Riley et al., 1983).

Greeno and his colleagues (e.g., Kintsch & Greeno, 1985; Riley & Greeno, 1988; Riley et al., 1983) and Briars and Larkin (1984) designed and implemented computer-simulation models of children's solutions to arithmetic word problems. In each model, problem solving is viewed as an interaction between conceptual knowledge and procedural knowledge. Specifically, for each model, the way children solve addition and subtraction word problems involves at least two steps. One step involves problem representation. It is assumed that children are able to construct representations corresponding to different types of problems. The other component common to all models is the selection of an appropriate strategy for solving the problem.

2. A schema-based model

The model developed by Riley et al. (1983), Kintsch and Greeno (1985), and Riley and Greeno (1988), henceforth referred to as the RGH model (because Riley et al. provided the initial presentation of the model), is a *schema-based model*. Successful problem-solving performance is accounted for by the availability of appropriate schemata in memory. As described previously, for the domain of arithmetic word problems, a schema is an organized structure consisting of elements and relations. The elements are organized in terms of quantitative (or numerical), temporal, and logical relations defining a general class of problems. The process of constructing a problem representation involves mapping the verbal problem statements with an appropriate assignment of specific quantities to the slots (or elements) of a relevant schema. Children are said to have, or be able to construct, schemata corresponding to different types of problems.

Greeno and his colleagues have identified three main classes of addition and subtraction word problems. Some problems involve a *change* schema. These problems describe situations in which an initial quantity is modified following a gain or loss of some amount, as illustrated in the following problem: Rick had 9 stickers. He gave 3 to Lisa. How many stickers does he have left? There are also problems that evoke a *combine* schema. These problems describe the combination of subsets into a superset, or the decomposition of a superset into subsets, as can be seen in the following example: There are 7 children on the playground. There are 3 boys among them. How many are girls? Finally, there are problems that involve a *comparison* schema. These problems describe situations in which two disjoint sets are being compared, as illustrated by the following problem: Rick has 6 stickers. Lisa has 3 stickers. How many stickers does Rick have more than Lisa?

There are other conceptual schemata proposed in the RGH model in addition to the change, combine, and compare schemata. An example is the *part-whole* schema, which is believed to be involved in some of the more difficult arithmetic word problems.

According to proponents of the RGH model, the conceptual schemata are activated during the initial representation of the problem. However, the successful solution of the problem is also contingent on the selection of appropriate *action* schemata. These action schemata are directly activated by the conceptual representation, although additional knowledge or processing may be required in order to select the appropriate schemata.

A key proposition of the RGH model is that "improvement in performance results mainly from improved understanding of certain conceptual relationships" (Riley et al., 1983; p. 154). In other words, the difficulty of arithmetic word problems is primarily a function of problem characteristics that affect problem representation. Indeed, on the basis of these conceptual relationships, or the semantic structure of the problems, word problems requiring simple addition and subtraction operations have been grouped into four major classes. Table 1 portrays each of these classes of problems and the specific subtypes within each class. The designations used in the table follow those in the classification systems of Carpenter and Moser (1982), Riley and Greeno (1988), Riley et al. (1983), and Willis and Fuson (1988).

Table 1. Classificaton of Addition and Subtraction Word Problems According to the RGH Model

---

CHANGE (CH)

Result Set Unknown:

1. Jack had 7 marbles.
   Then Rose gave him 2 more marbles.
   How many marbles does Jack have now?
2. Jack has 9 marbles.
   Then he gave 4 marbles to Rose.
   How many marbles does Jack have now?

Change Set Unknown:

3. Jack had 6 marbles.
   Then Rose gave him some more marbles.
   Now Jack has 10 marbles.
   How many marbles did Rose give him?
4. Jack had 8 marbles.
   Then he gave some marbles to Rose.
   Now Jack has 5 marbles.
   How many marbles did he give to Rose?

Start Set Unknown:

5. Jack had some marbles.
   Then Rose gave him 3 more marbles.
   Now Jack has 9 marbles.
   How many marbles did Jack have in the beginning?
6. Jack had some marbles.
   Then he gave 4 marbles to Rose.
   Now Jack has 5 marbles.
   How many marbles did he give to Rose?

COMBINE (CB)

Superset Unknown

1. Jack has 6 marbles.
   Rose has 3 marbles.
   How many marbles do they have altogether?
2. Jack and Rose have some marbles.
   Jack has 4 marbles.
   Rose has 2 marbles.
   How many marbles do they have altogether?

Subset Unknown:

3. Jack has 4 marbles.
   Rose has some marbles.
   They have 9 marbles altogether.
   How many marbles does Rose have?
4. Jack has some marbles.
   Rose has 5 marbles.
   They have 10 marbles altogether.
   How many marbles does Jack have?
5. Jack and Rose have 8 marbles altogether.
   Jack has 3 marbles.
   How many marbles does Rose have?
6. Jack and Rose have 7 marbles altogether.
   Jack has some marbles.
   Rose has 2 marbles.
   How many marbles does Jack have?

Table 1. (Continued)

COMPARE (CP)	
Difference Unknown:	
1. Jack has 9 marbles. Rose has 4 marbles. How many marbles does Jack have more than Rose?	2. Jack has 7 marbles. Rose has 4 marbles. How many marbles does Rose have less than Jack?
Compared Quantity Unknown:	
3. Jack has 3 marbles. Rose has 5 more marbles than Jack. How many marbles does Rose have?	4. Jack has 10 marbles. Rose has 4 marbles less than Jack. How many marbles does Rose have?
Referent Unknown:	
5. Jack has 8 marbles. He has 5 more marbles than Rose. How many marbles does Rose have?	6. Jack has 3 marbles. He has 6 marbles less than Rose. How many marbles does Rose have?
EQUALIZE (EQ)	
1. Jack has 6 marbles. If Rose wins 3 marbles she will have the same number as Jack. How many marbles does Rose have?	2. Jack has 9 marbles. If he loses 3 marbles he will have the same number as Rose. How many marbles does Rose have?
3. Jack has 8 marbles. Rose has 3 marbles. How many does Rose have to win to have the same number as Jack?	4. Jack has 9 marbles. Rose has 4 marbles. How many marbles does Jack have to lose to have the same number as Rose?
5. Jack has 2 marbles. If he wins 5 marbles, he will have the same number as Rose. How many marbles does Rose have?	6. Jack has 3 marbles. If Rose loses 5 marbles, she will have the same number as Jack. How many marbles does Rose have?

Change problems are composed of three subtypes, all of which involve an exchange of quantity. Whether the unknown is the start set, the change set, or the result set, is what determines the subtypes. Each subtype involves either addition or subtraction, yielding two broad classes of change problems. In change/join problems, which involve addition, there is an initial quantity (the start set) and a direct or implied action causes an increase (the change set) resulting in a new quantity (the result set). In change/separate problems, which involve subtraction, a subset (the change set) is removed from a given set (the start set), resulting in a new set (the result set). Note that in both classes of problems, the change occurs over time. Specifically, there is an initial condition at time t_1 which is followed by a change at time t_2 resulting in a final state at time t_3. As Table 1 shows, there are three change/join and three change\separate problems.

Both combine and compare problems involve static relationships for which there is no direct or implied action. Combine problems describe the relationship existing among a superset and its subsets. Two problem types exist. In the first type, the two subsets are given and a question is posed about the size of the superset. In the second type, one of the subsets and the superset are given and the problem consists in finding the size of the other subset (see Table 1).

Compare problems involve the comparison of two disjoint sets. It is customary to label one set the referent set and the other the compared set. The third set in these problems is the difference set, or the amount by which the larger set exceeds the smaller set. In compare problems, any one of the three sets can be the unknown. Furthermore, the larger set can be either the referent set or the compared set. These two factors make it possible to construct six different types of compare problems (see Table 1).

Equalize problems constitute the final class of addition and subtraction word problems. They are best conceptualized as a hybrid category of change and compare problems. Specifically, there is the same sort of action as in change problems; however, it is based on the comparison of two disjoint sets. As in compare problems, two disjoint sets are compared; and the question is asked about what could be done to one of the sets to make it equal to the other set. If the action to be performed is on the smaller of the two sets, then the problem is an equalize/join problem. On the other hand, if the action to be performed is on the larger set, then the problem becomes an equalize/separate problem. One can vary the unknown, as in compare problems, to produce three distinct problems of each of the two types of equalize problems (see Table 1).

Little research has been concerned with equalize problems. In addition, these problems are not frequently encountered in North American mathematics programs. It should be noted, however, that equalize problems are found in experimental mathematics programs developed in Japan and in the Soviet Union.

Several studies (e.g., Hiebert, 1982; Ibarra & Lindvall, 1980; Riley et al., 1983) have provided data on the relative difficulty of the various problem types across grade levels. Among the most difficult problems for children are those that involve the comparison schema, and those that involve the change schema with the starting set unknown. Up to the age of eight or nine years, children have a great deal of difficulty with these problems, and they tend to make characteristic errors. The distinctions between problems portrayed in Table 1 are also reflected in children's solution strategies. Children's solution processes tend to model the actions or relationships described in the problems.

The RGH model posits three levels of problem-solving competence in addition and subtraction word problems, from the least sophisticated (Level 1) to the most powerful (Level 3). Riley and Greeno (1988) describe the hierarchy as follows:

> *In general, Level 1 processes represent sets that are specified so models can be constructed directly as sentences are presented, and Level 1 processes count sets that are present in problem models. Level 2 processes represent sets that cannot be constructed externally, because they do not have specified numbers, and these processes represent relations between sets that are needed to make inferences. Level 3 processes can add part-whole relations to representations that already have other relations involving changes or comparisons of sets* (p. 62).

3. A model emphasizing problem-solving procedures

Another model of word problem solution has been proposed by Briars and Larkin (1984). This model, called CHIPS (Concrete Humanlike Inferential Problem Solver) by its proponents, solves arithmetic word problems concretely, doing the computer equivalent of laying out piles of poker chips to represent problem situations and then counting appropriate piles to find the answer. CHIPS also develops mathematical relations among the piles of chips.

According to Briars and Larkin, CHIPS has three levels of performance. These levels are viewed as models for stages in children's ability to understand arithmetic word problems. The simplest version of CHIPS represents problems using counters that have only one associated meaning tag. Thus, for instance, a chip can represent a penny that belongs to Erik. It cannot represent a penny that both belongs now to Eric and used to belong to his father. When CHIPS transfers the counter from the father's pile to Eric's, it loses the knowledge that the chip used to be in the father's pile. This version of CHIPS can solve simple problems like the following:

> *Jerry had three candies. He got five more. How many does he have now?*

To solve this problem, CHIPS makes a pile of three counters, representing Jerry's candies; then, adds to it a pile of five more counters; finally, counts the result. As is clear here, each counter is always understood as one of Jerry's candies. This version of CHIPS cannot solve the following problem:

> *Jerry had three candies. He got some more. Now he has eight. How many did he get?*

CHIPS makes a pile of three and adds counters to it until it contains eight. But then CHIPS cannot distinguish the original three counters from the additional counters; it cannot find the added counters in order to count them. The reason is that all the counters are identified with the single tag "Jerry's". This version of CHIPS cannot sustain a second identifier like "the new candies" or "the ones that Jerry got". This model has been called by its proponents CHIPS with single-role counters, or CHIPS_{SR} for short.

The second version of CHIPS has the ability to represent two roles for a counter. This CHIPS model solves the preceding problem easily. The second tag on each counter indicates whether it is an old or a new candy, and CHIPS readily counts Jerry's new candies. This model supports double-role counters and it has been referred to as CHIPS_{DR}. Note that CHIPS_{DR} can relate sets and subsets. By adding to the representation of each of Jerry's candies whether it is an old or a new candy, CHIPS represents the relation that Jerry's candies are a superset with two subsets. Of course,

CHIPS_{DR} does not have all the subtleties of relations between sets and subsets. For instance, this model of CHIPS is stumped by the following problem:

> *Sue had some baseball cards. She got five more. Now she has eight. How many baseball cards did she have before?*

The reason is that CHIPS concretely acts out each problem situation using chips. But it does not know how to use chips to represent the abstract idea "some". When simply following the basic language of the problem statement, CHIPS has no way to represent "some" as an indeterminate initial set to be affected by future actions.

The ability to represent "some" comes only with the third CHIPS model, which has several new abilities, all of which involve being able to re-represent the problem in a new way. This third model is called CHIPS with re-representation or CHIPS_{RR}. This model solves the preceding problem in one of the following ways: (i) It collects information from the problem in a knowledge structure called a time-reversal schema. It includes a first set (here unknown), a changed set, and a final set. It also includes the knowledge that, when convenient, time can be run backwards. Thus, to find the original number of baseball cards one can start with the final set, "undo" the change, and come up with the initial set. (ii) It collects information from the problem in another knowledge structure called a set-subset schema. Thus, it transforms the baseball card problem into a problem about a set of eight with a subset of five. The problem can then be solved by a variety of means, including separating five from eight.

In summary, the foregoing description of the three CHIPS models proposes that children's abilities in solving arithmetic word problems develop hierarchically, with each new component involving actions on entities of the previous levels. Specifically,

$CHIPS_{SR}$ represented only counters with a single meaning; $CHIPS_{DR}$ assigned chips two meanings and so could represent the set-subset relation; $CHIPS_{RR}$ could operate on a set-subset array to produce a new array.

CHIPS works directly on problems with actions that it can imitate, problems about getting, losing, finding, buying, etc. But there are arithmetic word problems that do not include actions. CHIPS needs additional knowledge to solve those problems. To illustrate, consider the following problems:

> *There are nine cookies on a plate. Five are vanilla and the rest are chocolate. How many are chocolate?*
>
> *There are seven dogs and three bones. How many more dogs are there than bones?*

CHIPS cannot solve these problems because there are no cues about what piles of chips to set out or to count. It needs additional knowledge that lets it interpret statements like the previous ones in order to lay out piles of chips. For the first problem, CHIPS sets out nine chips. It then responds to the phrase "Five are vanilla and the rest are..." as a cue to separate five chips from the rest. The question is then answered by counting the remaining chips (chocolate). In the second problem, CHIPS sets out two piles of chips representing dogs and bones. It then interprets the phrase "How many more..." as a cue to match dogs and bones and to separate dogs with bones from dogs without bones. The question is answered by counting the dogs without bones.

Notice that CHIPS is a minimal model, that is, it includes the smallest amount of knowledge that one could imagine being sufficient to solve the set of arithmetic word problems commonly encountered by children. For instance, CHIPS solves comparison problems not with any special extra mathematical knowledge, but with the same mechanisms for combining, separating, and counting chips used in other problems.

In sum, for Briars and Larkin (1984), several elements are believed to contribute to the degree of difficulty of word problems. The first element is whether the required action involved in the word problem is cued or not cued. Action-cued problems (e.g., all change problems) are easier to solve than problems with no action cues (e.g., all compare problems). Another aspect is whether the problem requires a single-role (sr) counter (e.g., Change 1 and Change 2) or a double-role (dr) counter (e.g., Change 3 and Change 4). A third difficulty factor hypothesized in this model involves re-representation: Word problems that need to be re-represented for solution are said to be more difficult than those not requiring this conceptual leap. All the aforementioned three elements can be viewed as aspects of conceptual knowledge.

In addition, Briars and Larkin (1984) observed that in compare problems there are additional language elements, such as consistent or conflicting language, which can affect problem difficulty. Obviously, this aspect relates to linguistic knowledge.

A comparison of the RGH and CHIPS models focusing on their explanatory power (i.e., children's performances they can explain) will quickly reveal that their predictions are the same concerning various empirical results. The hypothesized levels of competence are the same in both models. What is of prime importance in the present research is the emphasis both models place, implicitly or explicitly on conceptual knowledge.

C. Role of knowledge in children's solutions to arithmetic word problems

1. Knowledge and cognitive performance

Early information-processing research of human problem solving (e.g., Greeno, 1978; Newell & Simon, 1972) explored knowledge-lean problems, mostly puzzlelike problems such as the Tower of Hanoi problem. The focus was on basic information-processing capabilities that humans employ to solve problems in situations where they lack any specialized knowledge and skill. As a result, that knowledge offered limited insight into problem solving that requires domain-specific knowledge.

In contrast, more recent research in problem solving done in knowledge-rich domains such as physics, mathematics, and medicine, has repeatedly suggested that knowledge is at the heart of problem-solving performance. In a similar vein, most current theories of problem solving (e.g., Greeno, 1980; Simon, 1980) are based on the idea that solving a particular problem requires both domain-specific knowledge and knowledge of general strategies. They emphasize a new dimension of difference between individuals who show more or less ability in problem solving, namely, the possession of an accessible and usable body of knowledge.

As observed by Glaser (1984), the focus on the role of knowledge in cognitive tasks in general and problem-solving tasks in particular, rests on evidence from a variety of sources, including developmental studies of memory, reasoning, and problem solving (e.g., Carey, 1985; Chase & Simon, 1973; Chi, 1978; Chi & Koeske, 1983; Siegler & Klahr, 1982; Siegler & Richards, 1982), studies of expert and novice problem solvers (e.g., Chase & Simon, 1973; Chi, Glaser, & Rees, 1982; Larkin, McDermott, Simon,

& Simon, 1980), and process analyses of tests involving intelligence and aptitude items (e. g., Pellegrino & Glaser, 1982; Sternberg, 1977, 1981).

The conclusions drawn from the three sets of studies were unequivocal. The developmental studies of memory, reasoning, and problem solving clearly showed that changes in the knowledge base could produce sophisticated cognitive performance. With respect to memory, Chase and Simon (1973) and Chi (1978) found in their chess studies that high-knowledge subjects exhibited better memory performance than low-knowledge subjects. Concerning reasoning and problem solving, the acquisition of specific content knowledge as a determining factor in the emergence of increasingly sophisticated ability in those domains, was apparent in Carey's (1985) study of animistic thinking in children and in Siegler's rule assessment approach to the study of developmental change (e.g., Siegler & Klahr, 1982; Siegler & Richards, 1982).

An important conclusion reached by Carey was that the scant knowledge of 4- to 7-year-old children resulted in their inability to justify the inclusion and exclusion of humans, animals, plants, and inanimate objects under the concept 'alive'. The acquisition of domain-specific information about biological functions (brought about by school learning and world knowledge) allowed 10-year-olds to think about the properties of the concept and to reason appropriately.

In his version of the balance-scale task, Siegler showed that 5-year-olds could not solve some problems because of their failure to encode distance information in addition to weight information. Training these children to encode distance was to equip them with a more powerful "theory" about the relationship of weight and distance in balance-scale problems.

The studies of expert and novice problem solvers also converged on the notion that the problem-solving difficulty of novices could be attributed largely to the inadequacies of their knowledge bases and not to limitations in their processing skills (e.g., inability to use strategies for problem solving). In addition, it became apparent that the relation between the structure of the knowledge base and problem-solving processes was mediated by the quality of problem representation.

The third converging area of research, namely, process analyses of intelligence and aptitude test items, consisted of a set of empirical investigations of information-processing approaches to the study of intelligence and aptitude (e.g., Pellegrino & Glaser, 1982; Sternberg, 1977, 1981). In those investigations, researchers uncovered several interrelated components of performance differentiating individuals who scored high and those who scored low on intelligence and aptitude tests. Two such components involved conceptual knowledge, i.e., knowledge of relationships in the content domain, and procedural knowledge, i.e., knowledge of appropriate solution procedures in the domain. High-aptitude individuals appeared to display considerable skill in reasoning because of the high level of their knowledge of relationships in the content domain and because of their knowledge of procedural constraints of the problem domain. In other words, as individuals acquire knowledge, they become empowered to think and reason at a high level of sophistication.

2. Knowledge and children's mathematical cognition

As cognitive researchers became increasingly interested in studying learning and performance in specific domains of knowledge, mathematics has received a great deal of attention. A large portion of the research on mathematical thinking has been concerned with describing the knowledge, in the form of cognitive structures and processes, hypothesized to underlie competent performance on various mathematical tasks.

As observed by Putnam, Lampert, and Peterson (1990), the goal of this program of research (often dubbed the "knowledge-structure program") is to make explicit knowledge that is often implicit, but that is required for competent mathematical performance. Once this implicit knowledge is made explicit, it can be viewed as revealing the knowledge underlying understanding in the domain.

Two different approaches have been taken by cognitive researchers in building models of the knowledge underlying mathematical performance. The first approach involves carrying out a detailed analysis of children's performance on mathematical tasks. The second approach involves making a rich description of the mathematical content of various mathematical tasks.

Several domains of mathematics have been investigated from the perspective of the knowledge-structure program, including word problems, both in arithmetic (e.g., Carpenter & Moser, 1983; Riley et al., 1983; Vergnaud, 1982), and in algebra (e.g., Mayer, Larkin, & Kadane, 1984); problems involving rational numbers and fractions (e.g., Behr, Lesh, Post, & Silver, 1983); and problems involving decimal fractions (e.g., Resnick, Nesher, Leonard, Magone, Omanson, & Peled, 1989)

Types of knowledge hypothesized to underlie mathematical competence: The distinction between knowledge and understanding of the concepts of mathematics on the one hand, and knowledge of the procedures of mathematics on the other, has been the subject of much debate over the years. For instance, almost a century ago, McLellan and Dewey (1895) argued for understanding and presented their optimal mathematics curriculum intended to raise the level of comprehension beyond that prevailing in classrooms at the time. In contrast, Thorndike (1922) presented the case for skill learning (i.e., drill by practice) and described in detail how skills should be taught to maximize retention. Brownell (1928) opposed the emphasis on learning isolated skills and argued forcefully for an increased emphasis on understanding. In essence, the same debate continued through Gagné's (1977) emphasis on skill learning versus Bruner's (1960) case for understanding. The modern version of the debate involves the distinction between conceptual knowledge and procedural knowledge in mathematical problem solving. Currently, many cognitive researchers and mathematics educators (e.g., Hiebert, 1985) consider the distinction to be useful for thinking about performance on mathematical tasks. Certainly, researchers vary in their definitions of the two types of knowledge. However, for the purpose of this thesis, the following two definitions are adopted. Conceptual knowledge is knowledge of relationships in a problem domain, or as aptly described by Hiebert and Lefevre (1985), "a connected web of knowledge, a network in which the linking relationships are as prominent as the discrete pieces of information" (pp. 3-4). Procedural knowledge is knowledge of principles, rules, algorithms, and procedures used to solve mathematical tasks.

As previously observed, many researchers have directed their efforts at describing the knowledge underlying addition and subtraction word problems. An important distinction in that literature is that between conceptual knowledge and linguistic knowledge.

There are investigators who have proposed that children's improved solution performance on arithmetic word problems could be attributed to the development of logico-mathematical knowledge, which is the type of conceptual knowledge relevant to the domain of arithmetic word problems. As is apparent in the discussion of the RGH and CHIPS models, Riley and Greeno (1988), Riley et al. (1983), Briars and Larkin (1988), and Resnick (1983) were proponents of the logico-mathematical view. For instance, Greeno and his colleagues have explained poor performance on some arithmetic word problems (e.g., Change 5 and Change 6 problems) by invoking children's lack of sufficient knowledge concerning part-whole set relations. They have, in effect, stated that children fail these problems because they have not yet acquired the part-whole schema, a key conceptual achievement in their problem-solving model. Resnick (1983) was unequivocal about the crucial role of the part-whole schema:

> *With the application of a part-whole schema to quantity, it becomes possible for children to think about numbers as compositions of other numbers. This enrichment of number understanding permits forms of mathematical problem solving and interpretation that are not available to younger children.* (p. 114)

It is not surprising then that Riley and Greeno (1988) have made the part-whole schema the hallmark of their most advanced level of skill in solving arithmetic word problems. Briars and Larkin (1988) also have recognized the importance of the part-whole schema. They thought one could only propose additional ways of solving some of the problems.

Recall that Briars and Larkin have emphasized that children's understanding of double-role counters and of re-representation were of prime importance to their acquisition of more advanced problem-solving skills. Clearly, all these are aspects of conceptual or logico-mathematical knowledge.

There are other researchers (e.g., Cummins, 1991; Cummins et al., 1988) who have attributed the improvement in children's skills in solving arithmetic word problems to the development of language-comprehension skills. According to them, the words used in some problems are too difficult for children to understand, even for children who possess the required logico-mathematical knowledge.

A case in point is the use of comparative verbal forms (e. g., more than, less than, fewer than) in problem texts. Note that these forms are always present in the usual formulations of all compare problems. For Cummins et al. (1988), the view that language difficulties underlie poor solution performance has some empirical support in results showing that children often transform comparatives into simple assignment or possession terms when they are asked to retell word problems (Cummins et al., 1988), skip over them when reading phrases containing them (De Corte & Verschaffel, 1986), and perform better when problems are reworded to exclude the comparative terms (e.g., De Corte & Verschaffel, 1985; Hudson, 1983).

The major thrust of the present research was to investigate further the role of conceptual and linguistic knowledge in children's performance on arithmetic word problems. Most accounts of the development of children's skills in solving these stress the role of the understanding of part-whole set relations or the part-whole schema.

Empirical research on the part-whole schema: To date, the part-whole schema has been the single most important piece of conceptual knowledge proposed by researchers espousing the logico-mathematical view to explain the development of competence in solving arithmetic word problems. However, empirical evidence for its critical role is both scant and conflicting.

Wolters (1983) developed an instructional sequence of 26 to 30 lessons that provided diagrams representing part-whole relations. This instructional sequence was used in third- and fourth-grade classrooms over a period of two to three months. Instruction involved relating diagrams to addition and subtraction formulas and drawing part-whole diagrams for some word problems. Following the instruction, children solved two-step problems in the combine category more successfully than did children in a control group; however, their performance was poorer than that of the control group in the change and compare categories. Hence, a training program stressing the general schema for making inferences was not quite effective in Wolters' study.

Using a developmental sequence perspective, Dean and Malik (1986) asked children to solve problems in the change category, and to anticipate questions when the given information in the problems had been presented. They found that children who could not solve at least one problem among the Change 5 and Change 6 problems were unable to anticipate questions. Dean and Malik concluded that knowledge of the part-whole schema is a pre-requisite for acquiring the change schema. What this implies is that to solve any change problem, children must have developed the part-whole schema. Clearly, this conclusion was based on the assumption that solutions of Change 5 and Change 6 problems involve using the part-whole schema, and the assumption that anticipation of questions involves using the change schema. Riley and Greeno (1988), however, have advanced an alternative interpretation for Dean and Malik's results.

Willis and Fuson (1988) taught two classes of second graders of average and above-average mathematical ability to use differing schematic drawings to represent differing categories of addition and subtraction word problems. Children were asked to enter the numbers used in the problems onto the schematic drawings and then to use the drawings to facilitate the choice of the solution procedure. Clearly, this approach was emphasizing the semantic structure of the problem situation: Children were observed to choose a schematic drawing that matched the structure of a given word problem, filling the numbers given in the word problem into appropriate locations in the schematic drawing. Use of the part-whole schema was indicated by use of a Put-Together drawing consisting of two parts and a whole.

The investigators found little evidence in support of Resnick's (1983) suggestion that children use a part-whole schema (like the Put-Together drawing) to solve all types of arithmetic word problems. There was also little support for the more restricted proposal of Riley and Greeno (1988), Riley et al. (1983), and Kintsch and Greeno (1985) that children solve some of the more difficult problems (e.g., Change 5 and Change 6 problems) by using a part-whole schema.

Fuson and Willis (1989) used a more complete testing procedure addressing more adequately than in their earlier study issues of problem difficulty and choices of alternative drawings. This permitted a better test of the hypothesis concerning children's use of the part-whole schema to solve addition and subtraction word problems. Their findings revealed that children did not use the part-whole schema to solve the most difficult change and compare problems.

Morales et al. (1985) asked a group of third grade children and a group of fifth and sixth grade children to solve problems and to sort problems into different categories. Their results for problem solving were generally consistent with findings reported by Riley et al. (1983). Their cluster analysis of the sorting data revealed that both groups sorted change problems into a category, with subtypes based on whether the unknown was the result set, the change set, or the start set. Fifth and sixth grade children also formed a category containing two combine problems along with Compare 3, Compare 4, Compare 5, and Compare 6 problems. Riley and Greeno (1988) interpreted these results as supporting the proposal that "difficult compare problems are understood by older children in a way that includes part-whole relations involving subsets of the larger of the two sets in the problem" (p. 55). Clearly, however, these results are only suggestive.

F. E. Fischer (1990) taught 42 kindergarten children using a "part-part-whole" curriculum that stressed set-subset relationships. These children were later found to be more successful in solving addition and subtraction word problems, even though addition and subtraction applications were not part of the instructional treatment.

Two unpublished studies (Rathmell, 1986; Tamburino, 1982) were also cited by Riley and Greeno (1988) as supportive of the proposal that the part-whole schema is involved in performance on arithmetic word problems. The apparent success of Rathmell's and Tamburino's instructional sequences was due, according to Riley and Greeno, to their use of specific problems of various types mapped onto representations of part-whole relations among quantities, as opposed to the procedures by Wolters (1983).

In summary, the evidence is at best mixed that the part-whole schema is a component of successful performance on arithmetic word problems, which means that more research is needed about the question.

Concerning the importance of other aspects of conceptual knowledge, that is those proposed in the CHIPS model (e.g., understanding of double-role counters, re-representation), evidence beyond that reported by Briars and Larkin (1988) is lacking.

The position taken in the present research is that a different perspective or framework may be more productive for research dealing with the role of conceptual knowledge in children's performance on arithmetic word problems. This is what is attempted in this research.

D. Latent trait theory and a model for component processes

1. Latent trait theory

Overview of latent trait theory

Latent trait theory, also known as item characteristic curve theory or item-response theory (IRT) is currently receiving considerable attention from measurement experts and testing practitioners. This popularity of IRT can be accounted for by its potential for providing viable answers to many measurement and testing problems.

In many respects, the *actual* development of IRT is due to the independent work of Lord (1952) and Rasch (1960). Much of Lord's earlier work was of theoretical interest only, because of the mathematical intractability of his proposed formulations. With Birnbaum's (1968) development of logistic models, the applications of Lord's theoretical works became possible. However, it was only with the development of powerful computers and sophisticated software that applying IRT to real measurement and research problems became practicable.

At the core of IRT is the item characteristic curve (ICC), which is a description of the relationship between an examinee's ability level on the trait (or construct) being measured by the item and the probability that the examinee will respond to the item correctly. In IRT, the ability level of an examinee on the trait being measured is denoted by θ and is simply referred to as ability. The probability that examinee j responds correctly to item i in a test is denoted by $P(X_{ij} = 1/\theta_j, \beta_i)$, a notation sometimes shortened to $P_i(\theta)$. Both expressions are used in this thesis.

The basic idea of IRT is that, if the relationship between θ_j and $P(X_{ij} = 1)$ is known for each item in a test, the item characteristics of each item, the ability of each examinee, and the measurement error associated with each score can be derived mathematically. Furthermore, the characteristics of any test made up of items with known item characteristics can also be derived. The obvious payoff is that knowledge of the characteristics of each item will lead to better solutions for various measurement and testing problems, such as problems in adaptive testing and equating.

One important characteristic of IRT is its item-level orientation. More specifically, IRT models make definite statements about the relationship between the probability of answering an item correctly and the examinee's ability or level of achievement.

However, as noted by Bejar (1983), a model that considers groups, such as classrooms or schools, has been proposed by Bock and Mislevy (1981). Even this group-level IRT model has retained the item-level orientation of IRT. Indeed, for many psychometricians the power of IRT resides precisely in the willingness to make a statement about performance at the item level.

Assumptions and models in IRT

As a test theory based on a family of mathematical models, there are certain assumptions that must be made about test scores, and these are conditions under which the model under consideration can be assumed to hold. IRT makes strong assumptions about test data. There are three basic assumptions that must be made when IRT is used to model examinees' test behavior.

The first assumption involves the dimensionality of the latent space. In general, IRT models assume that the probability of a correct response by an examinee can be attributed to his or her standing on a specific number (k) of latent traits or abilities. Geometrically, an individual's position on each of the latent traits can be conceptualized as a point in k-dimensional space. For most applications of IRT, it is assumed that $k = 1$, i.e., that the latent space is unidimensional, which implies that examinee performance can be accounted for by a single latent trait.

Although most IRT models currently in use involve unidimensional latent spaces, significant progress has been made in recent years in the development and applications of multidimensional IRT (MIRT) models (e.g., Batley & Boss, 1993; Carlson, 1987; McKinley, 1989; Reckase & McKinley, 1985). The development of MIRT models originated in the belief that most test data are multidimensional in nature. Many psychometricians (e.g., Ackerman, 1989; Ansley & Forsyth, 1985; Harrison, 1986; Humphreys, 1962, 1986; Reckase, 1979, 1985; Reckase & McKinley, 1991; Traub, 1983) have emphasized that real test data most often cannot be well modeled by unidimensional models. They consider unidimensionality as a "fragile assumption" (Traub, 1983; p. 59).

Note that the above discussion of the dimensionality of the latent space underlying test scores does not make a distinction between major and minor dimensions. This is in line with the traditional IRT definition of dimensionality. However, in recent years, Stout and his colleagues (e.g., Nandakumar, 1991, 1993; Nandakumar & Stout, 1993; Stout, 1987, 1990) have proposed a new conceptualization of test dimensionality. They make a fundamental distinction between *traditional* dimensionality and *essential* dimensionality. Whereas traditional dimensionality refers to the number of dimensions in the latent space, essential dimensionality refers to the number of dominant dimensions only. Clearly, it may be important in some situations to be able to assess the number of dominant dimensions underlying a set of test scores.

The second assumption relates to conditional or local independence, which means that item scores are statistically independent for examinees at the same ability level. For this assumption to be satisfied, an examinee's performance on an item must not affect his or her answers to any other items on the test. Consequently, this assumption is violated if, for instance, the content of an early item in a test provides clues to the correct response for a later item. When local independence holds, the probability of occurrence of any pattern of item scores for an examinee is simply the product of the probability of individual item response probabilities. To illustrate, in a situation in which local independence is satisfied, the probability of the occurrence of the five-response pattern V = (0, 1, 1, 0, 1), where 1 denotes a correct answer and 0 an incorrect answer, is given by $(1-P_1)$ (P_2) (P_3) $(1-P_4)$ (P_5), where P_i is the probability that the examinee responds correctly to item i and $1-P_i$ is the probability that he or she responds incorrectly to the item. Conversely, the assumption of local independence is satisfied when the probability of any response pattern for an examinee is equal to the product of the probabilities associated with the examinee's responses to the items. If a set of test items measures a single ability, the assumption of local independence is necessarily satisfied; that is, local independence is a consequence of unidimensionality. In other words, the assumption of local independence in a case in which a single ability dimension accounts for item performance and the assumption of a unidimensional latent space are equivalent. It is important to realize that the assumption of local independence does not imply that test items are uncorrelated over the total group of examinees. Positive correlations between pairs of items will result whenever there is variation among the examinees on the ability measured by the test items. Item scores are uncorrelated only for examinees at a given ability level.

The relationship between test dimensionality and local independence has been formalized by McDonald (1981, 1982), who has argued that a meaningful definition of dimensionality should be based on the principle of local independence. For him, a set of test items is k-dimensional if, for examinees with an identical profile of ability scores in a k-dimensional space, the covariation between items in the set is equal to zero. He defines the dimensionality of a set of test items as the number of traits needed to satisfy the assumption of local independence.

Stout (1987, 1990) has provided a weaker form of local independence that he has referred to as essential independence from which flows his theory of essential dimensionality. Essential independence requires only that, in an N-item test, as N goes to infinity, the average absolute covariation over item pairs is small in magnitude for all values of θ.

The third assumption of IRT concerns the assumed shape of the item characteristic curve (ICC). An ICC is a mathematical function that relates an examinee's probability of success on an item to the ability measured by the set of items in the test. More specifically, it is the nonlinear function for the regression of item scores on the trait or ability measured by the test. The main difference among currently popular IRT models is in the mathematical form of the ICC. The test developer or researcher has to choose one of the many mathematical functions that could serve as ICCs. In doing so, some assumptions are being made that can be evaluated later by how well the chosen model accounts for the test data. If the model accounts for the data, the appropriateness of all the model assumptions is justified. One is forced to conclude that one or more of the assumptions of the model are untenable when the fit of the model to test data is unsatisfactory.

Some of the current IRT models are based on the form of the normal ogive. This function is approximated by the logistic ogive, which is less mathematically cumbersome.

In general, with the common IRT models, the number of parameters in the mathematical function determines the form of the ICC, resulting in what has been generally referred to as an n-parameter logistic model, where usually, n = 1, 2, or 3.

In the currently popular unidimensional IRT models, invariably, there is a parameter for location (difficulty, β_i); some models include a parameter for strength of relation to the latent trait (discrimination, a_i); and some models include a parameter for lower asymptote (guessing or pseudo-guessing, c_i).

The difficulty parameter β_i corresponds to the point of inflection on the ability (θ_j) scale. For the 1-parameter and 2-parameter models, this is the point on the ability scale at which an examinee has a 50% chance of correctly answering an item. For the 3-parameter model, this the point at which the probability of correctly answering an item is $(1 + c_i)/2$, where c_i is the lower asymptote parameter.

Theoretically, difficulty values can range from -infinity to +infinity; in practice, values usually are in the range -3 to +3, when θ_j is scaled to have a mean of 0 and a standard deviation of 1.0. Similarly, examinee ability values can range from -infinity to +infinity, but values beyond -3 or +3 seldom are seen. Items with high values of β_i are hard items, with low-ability examinees having low probabilities of correctly responding to the item. Items with low values of β_i are easy items, with most examinees, even those with low-ability values, having at least a moderate probability of getting the item right.

The 2- and 3-parameter models have a discrimination parameter (a_i) that allows items to differentially discriminate among examinees. Technically, a_i is defined as the slope of the ICC at the point of inflection. This parameter can range in value from -

infinity to +infinity, with typical values being less than or equal to 2.0 for multiple-choice items. The higher the value of the discrimination parameter, the more sharply the item discriminates at the point of inflection.

The 3-parameter model also has a lower asymptote parameter (c_i), which is sometimes referred as a guessing or pseudo-guessing parameter. This parameter allows for examinees, even those with low ability, to have perhaps substantial probability of correctly answering even moderate or hard items. Theoretically, c_i ranges from 0.0 to 1.0., but is less than 0.3 in most cases.

The 1-, 2-, and 3-parameter models are now described in more detail. The 2-parameter logistic model is presented first, as the other two models are easily derived from it.

The 2-parameter logistic model: The 2-parameter logistic (2-PL) model has been proposed by Birnbaum (1968) as a more mathematically tractable alternative to Lord's (1952) 2-parameter normal-ogive model. The 2-PL model is defined mathematically as:

$$P(X_{ij}=1/\theta_j,\ \beta_i) = \frac{\exp[Da_i\ (\theta_j\ -\ \beta_i)]}{1\ +\ \exp[Da_i\ (\theta\ -\ \beta_i)]} \qquad [1]$$

where a_i and β_i are respectively the discrimination and difficulty parameters as defined above, and exp(x) is used to denote the constant e raised to the power x. The constant D is an arbitrary scaling factor. It is customary to set $D = 1.7$, since it has been shown that when $D = 1.7$, values of $P(X_{ij} = 1/\theta_j, \beta_i)$ for the 2-PL and the normal-ogive models differ, in absolute value, by no more than .01 for all levels of θ_j (Lord & Novick, 1968).

Hence, by setting D = 1.7, the interpretation of item parameters in the logistic models is the same as in the normal-ogive models.

Note that there is an alternative way of writing the equation for the 2-PL model. By dividing the numerator and denominator of Equation 1 by $\exp[Da_i\,(\theta_j - \beta_i\,)]$, then $P_i\,(\theta)$ becomes:

$$P(X_{ij}=1/\theta_j,\ \beta_i) = \frac{1}{1 + \exp[-Da_i\,(\theta_j - \beta_i)]} \qquad [2]$$

which is also written as:

$$P(X_{ij}=1/\theta_j,\ \beta_i) = \{1 + \exp[-Da_i\,(\theta_j - \beta_i)]\}^{-1} \qquad [3]$$

Four sample ICCs for the 2-PL model are presented in Figure 1. As can be seen, the curves are not parallel; they have different slopes, which reflects the fact that the discrimination parameter values are different. Also, all three curves have lower asymptote values of zero, indicating that in the 2-PL model it is assumed that examinees do not guess when attempting to solve items.

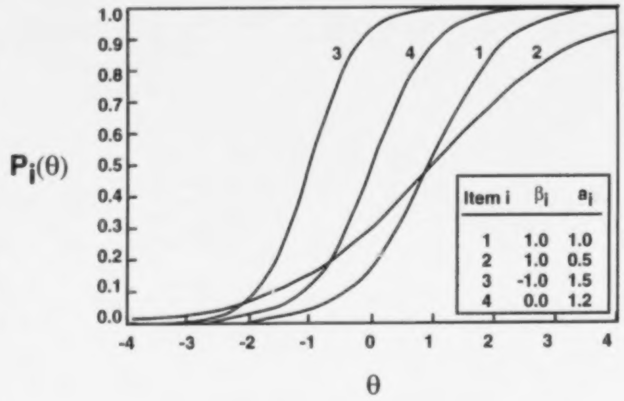

Figure 1. 2-PL ICCs for Four Items

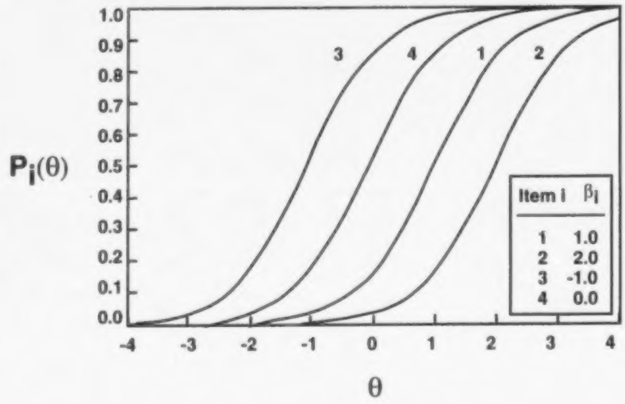

Figure 2. 1-PL ICCs for Four Items

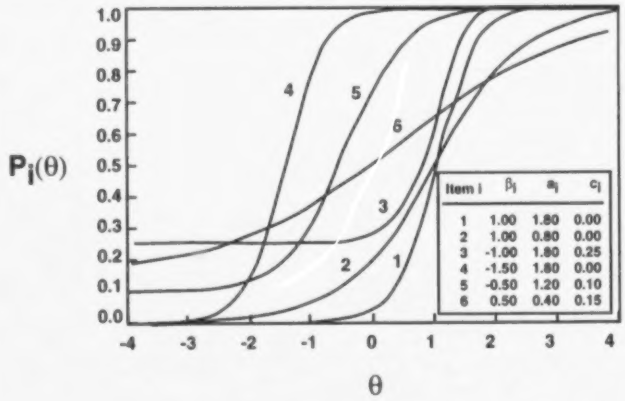

Figure 3. 3-PL ICCs for Six Items

The 1-parameter logistic model: The 1-parameter logistic (1-PL) model is a special case of the 2-PL model in that it assumes that all items have equal discriminating power and vary only in difficulty. Thus, for this model, the discrimination parameter is represented by a constant, a, instead of the variable a_i. The 1-PL model is mathematically defined as:

$$P(X_{ij}=1/\theta_j,\ \beta_i) = \frac{\exp[Da\ (\theta_j - \beta_i]}{1 + \exp[Da\ (\theta_j - \beta_i)]} \qquad [4]$$

If a new scaling is adopted such that $\theta_j^* = Da\theta_j$ and $\beta^*_i = Da\beta_i$, then the 1-PL model can also be written as:

$$P(X_{ij}=1/\theta_j^*,\ \beta_i^*) = \frac{\exp(\theta_j^* - \beta_i^*)}{1 + \exp(\theta_j^* - \beta_i^*)} \qquad [5]$$

From Equation 5, it can be clearly seen that with the 1-PL model, the probability that an examinee responds correctly to an item is a function of the examinee's ability and the difficulty of the item only. This model is the most restrictive of all three logistic models because of the two constraints imposed on the discrimination parameter (i. e., $a_i = a$ for any item i) and the pseudo-guessing parameter (i. e., $c_i = 0$ for any item i). The 1-PL model is also known as the RASCH model. George Rasch was a Danish mathematician who developed a model equivalent to the 1-PL model.

In Figure 2 are shown four sample ICCs for the 1-PL model. Except for an axis translation for representing their different locations on the θ scale, the curves are the same, which is expressed by their parallelism. This makes sense because for the 1-PL model, it is assumed that only the difficulty parameter influences performance on the items. In addition, as in the ICCs presented in Figure 1 for the 2-PL model, the lower asymptote values of all four curves are zero.

The 3-parameter logistic test model: The 3-parameter logistic (3-PL) model is constructed from the 2-PL model by adding a third parameter, denoted c_i. The mathematical representation of the model is as follows:

$$P(X_{ij}=1/\theta_j, \beta_i) = C_i + (1-C_i)\frac{\exp[Da_i(\theta_j - \beta_i)]}{1 + \exp[Da_i(\theta_j - \beta_i)]} \qquad [6]$$

The parameter c_i is the lower asymptote of the ICC and it gives the probability that low ability examinees respond correctly to an item. This parameter is included in the model to "correct" the misfit of ICCs at the lower end of the ability continuum where guessing can be a factor in performance on objective tests. However, typically, c_i is ascribed values that are smaller than the values that would result if examinees of low ability were to randomly guess in responding to items. The 3-PL model is the least restrictive of the three logistic models described here since it involves more parameters to be estimated. In other words, the requirements that ai and ci be equal to specific constants, as in the 1-PL model, are removed in the 3-PL model.

Six sample ICCs for the 3-PL model along with their respective parameter values are displayed in Figure 3. Since this model contains all three parameters, it can be used to highlight the role of the parameters in shaping ICCs. For example, in Figure 3, a comparison of Items 1, 2, and 3 with Items 4, 5, and 6 makes clear the influence of β_i (difficulty parameter) on the location of ICCs. Items 1, 2, and 3, which are more difficult, are located at the higher end of the θ (ability) scale, whereas Items 4, 5, and 6, which are easier are shifted to the lower end. Likewise, a comparison of Items 1, 3, and 4 with Items 2, 5, and 6 highlights the effect of a_i on the steepness of ICCs. Finally, a comparison of Items 1, 2, and 4 with Items 3, 5, and 6 informs about the role of c_i (lower asymptote paramneter) in shaping ICCs.

The influence of the difficulty, discrimination, and lower asymptote parameters on the shape of ICCs is more clearly illustrated, respectively, in Figures 4, 5, and 6. In each of these figures, only the values of the parameter whose effect is being shown differ for the three items under consideration.

Note that the 1-PL, 2-PL, and 3-PL models are applicable to binary items, that is, test items that are scored dichotomously. There are other models suited for other scoring formats, such as the nominal response model (Bock, 1972; Samejima, 1969), the graded response model (Samejima, 1969), and the continuous response model (Samejima, 1973). Thissen and Steinberg (1986) provided a useful taxonomy of item response models in which they organized models as members of three distinct classes, within which the models are distinguished only by assumptions and constraints on their parameters.

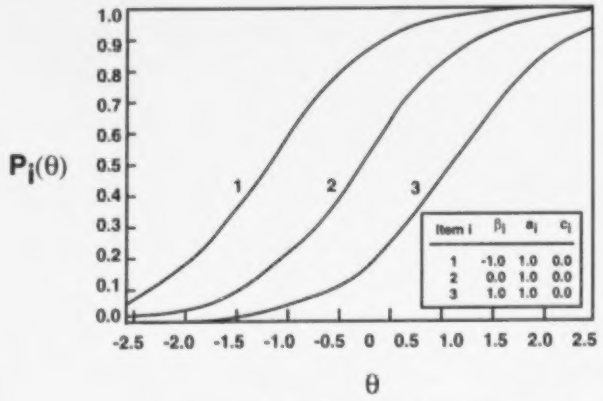

Figure 4. 3-PL ICCs for Three Items
(same a and c values, different β values)

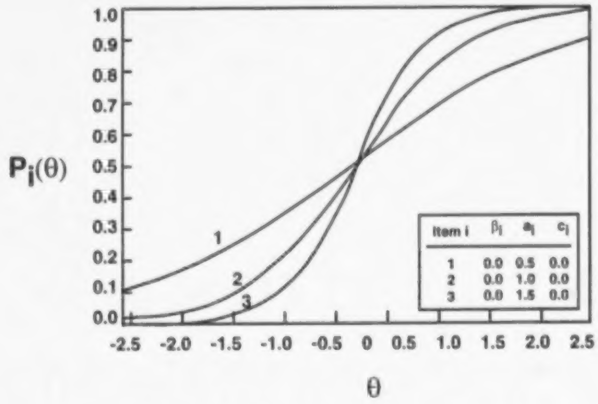

Figure 5. 3-PL ICCs for Three Items
(same β and c values, different a values)

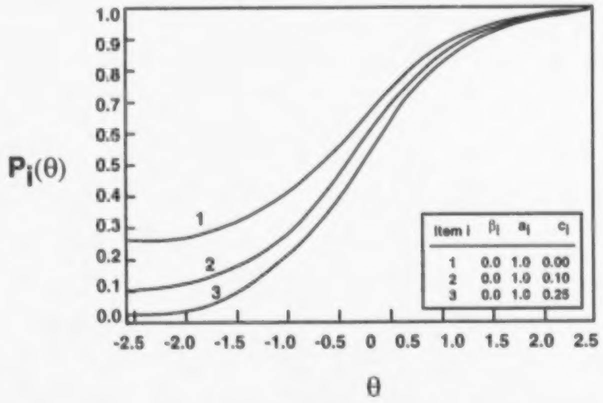

Figure 6. 3-PL ICCs for Three Items
(same a and β values, different c values)

Estimation of ability and item parameters

For IRT to be used in solving measurement problems, procedures must be implemented for estimating the parameters of the items and of the examinees. The various parameter-estimation methods being used in practical applications of and research with IRT models include conditional maximum likelihood, joint maximum likelihood, marginal maximum likelihood, Bayesian, and heuristic (or intuitive) methods.

In practice, two main estimation situations arise: (i) estimation of ability parameters when item parameters are known, that is, when the item-parameter estimates obtained from an earlier calibration are taken to be the true values; and (ii) simultaneous estimation of ability and item parameters. For the first situation, conditional maximum likelihood is the most frequently used method. For a more detailed presentation of parameter-estimation methods, see for example, Baker (1987) and Hambleton (1989).

The parameter-estimation methods are based on asymptotic statistical theory. As a result, to optimally estimate parameters in IRT, large data sets are required. In general, suitable test lengths and sample sizes are needed. However, it is difficult to state an omnibus rule which is satisfactory in every situation, as many factors are involved. One such factor is the IRT model being used: the greater the number of model parameters, the larger the sample size needs to be. For instance, with N examinees, n items, and the 3-PL model, there are N + 3n parameters to be estimated; for the 2-PL model, there are N + 2n parameters; and for the 1-PL model, there are N + n parameters to be estimated.

Another factor is the choice of parameter-estimation method. For example, it is now known that Bayesian methods give better estimates than maximum likelihood methods when both N and n are small. A third factor to consider is the distribution of ability in the sample of examinees. More specifically, other things being equal, heterogeneous samples are more satisfactory than homogeneous samples. A fourth factor is whether estimation involves only item parameters or ability parameters or both. In this regard, larger numbers of examinees and items are needed when both ability and item parameters are being estimated. Finally, when IRT is being used for important applications, only small errors are acceptable, making it crucial to have larger examinee samples and longer tests.

Various computer programs that implement the aforementioned estimation methods are now available for parameter estimation in IRT. Several of the most widely used include BICAL (Wright & Stone, 1979), BILOG (Mislevy & Bock, 1984), LOGIST (Wingersky, 1983; Wingersky, Barton, & Lord, 1982), MULTILOG (Thissen, 1986), ASCAL (Assessment Systems Corporation, 1988), RASCAL (Assessment Systems Corporation, 1988), NOHARM (Fraser, 1981; Fraser & McDonald, 1988), ANCILLES (Urry, 1978), and OGIVIA (Urry, 1977). For more information on these programs, see for example, Hsu and Yu (1989), Hambleton (1989), Mislevy and Stocking (1989), and Yen (1987).

Assessment of fit of models to test data

Among the advantages of IRT models is the fact that they are falsifiable. This is so because "it is possible to make various tangible predictions from the model and then to check with observed data to see if these predictions are approximately correct" (Lord, 1980; p. 15). IRT theorists and researchers have directed considerable effort toward the goal of developing and refining goodness-of-fit tests by which to falsify IRT models in specific situations. However, it is important here to realize that "there exists no basis in inductive logic for concluding that a model fits data, yet it is just this conclusion we usually wish to justify" (Traub, 1983; p. 57). What accounts for this fact is that tests of fit can only provide means for rejecting hypotheses, not accepting them. In other words, goodness-of-fit investigations are really about the assessment of model-data misfit.

IRT has the potential for contributing to the solution of many measurement problems arising both in research and applied contexts. However, the success of specific applications is not assured merely by using IRT models; that is, simply by processing test data through one of the computer programs for the estimation of IRT parameters. IRT model advantages can be obtained only when there is a satisfactory fit between the model and the test data on hand. An IRT model showing clear misfit for a specific set of test scores will not reveal any measurement properties of these scores. This means that goodness-of-fit investigations are critical for any careful IRT modeling endeavor.

There are three generally accepted strategies taken by researchers when investigating the appropriateness of an IRT model to test data. First, one can determine if the test data satisfy the assumption of the model of interest. Recall that the adoption of a given model implies specific assumptions about the test data. In the case of IRT, these assumptions are strong assumptions. For instance, adoption of the 1-PL model implies the following assumptions about the test data: (i) that only one ability is measured by the test; (ii) that all items on the test have the same discriminating power; and (iii) that guessing is not a factor in test performance, even for low-ability examinees.

Second, one can determine if the expected advantages derived from the use of the IRT model are obtained. Three advantages are obtained when an IRT model fits a set of test data: (i) Examinees' ability estimates are obtained on the same ability scale and can be compared even though examinees may have taken different sets of items from an item pool measuring the ability under consideration. In other words, ability estimates are invariant; that is, they do not vary as a function of the specific items taken from a given item pool. This model feature is referred to as test-free person measurement; (ii) Item statistics are obtained which are invariant, that is, they are not dependent on the sample of examinees used in the estimation of item parameters. This model advantage is called person-free item calibration; and (iii) Information is provided about the precision of ability estimates at each point on the ability scales. This is accomplished by conditional standard errors, that is, standard errors that are dependent on ability level. Attempts to check model features can be based on features described in (i) and (ii) above.

Third, one can determine the closeness of fit between the predictions derived from an IRT model and actual or simulated outcomes. More specifically, one substitutes estimated parameters for true parameters and then one can see whether or not an approximate fit is obtained between predictions and outcomes. Thus, the most obvious indicator of (mis)fit is the residual difference between the observed item score for an examinee and his or her expected item score under the IRT model under consideration.

The IRT literature is replete with discussions of methods for investigating goodness of fit in IRT modeling. For instance, Traub and Lam (1985) list the most widely used indicators of (mis)fit; Hambleton (1989), Hambleton and Murray (1983), Hambleton and Swaminathan (1985), Hambleton, Swaminathan, and Rogers (1991) review the possible methods for checking model assumptions, expected model features, and model predictions of real and simulated test data; and Hattie (1984, 1985) evaluates numerous methods for assessing the unidimensionality of tests and items.

Latent trait theory has had a substantial impact on testing research and practice, especially in North America. However, as useful as they are for constructing tests and analyzing test data, standard IRT models do not address the cognitive processes that examinees must use to have a high probability of making a correct response. They also do not address the features that make some tasks more difficult than others. Recall that the two main model parameters only indicate the relative proficiencies of examinees (θ) and the relative difficulties of tasks (β_i).

To fill this gap, IRT models needed to be extended to include, for example, parameters reflecting the processing requirements of tasks or test items. Many researchers in Europe as well as in North America were aware of the limitations of standard IRT, but those in Europe were among the first to propose extending standard IRT to make it useful for investigating problem-solving processes. According to Fischer and Forman (1982), "latent trait theory in Europe has developed more towards a theory of psychological measurement and towards modeling the process of problem solving, . . ." (p. 397; emphasis added).

In that context, as noted by Fischer and Forman, the Rasch model has been the most influential model. In effect, two extensions of the Rasch model have been particularly important and are "currently attracting the attention of cognitive psychologists and psychometricians" (Hambleton, 1989; p. 159). The first extension has been proposed by Fischer and his colleagues (e.g., Fischer, 1973, 1983; Fischer & Forman, 1982; Spada & McGaw, 1985) and has been referred to as the *linear logistic test model (LLTM)*. The second extension is found in the work of Embretson and her colleagues (e.g., Embretson, 1984, 1985; Embretson, Schneider, & Roth, 1986; Whitely, 1980) and has been called the *multicomponent latent trait model (MLTM)*. Both the LLTM and the MLTM are constrained Rasch models. However, whereas the LLTM is a unidimensional model, the MLTM is a multidimensional model. The present research was concerned with modeling applications of the LLTM.

2. The linear logistic test model

Description of the model

The LLTM was developed in the early 1970s (G. H. Fischer, 1973), but until recently, it was virtually unknown in North American psychometrics. As stated above, it is an extension of the Rasch model. As is well known, the Rasch model predicts the encounter of person j with item i by just two parameters, the person's ability, θ_j , and the difficulty of the item, β_i. The probability that the encounter leads to a correct solution of the item , $P(X_{ij} = 1/\theta_j, \beta_i)$, is modeled as follows:

$$P(X_{ij}=1/\theta_j, \beta_i) = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)} \qquad [7]$$

Note that Equation 7 is equivalent to both Equations 4 and 5. In other words, Equation 7 is an alternative way of mathematically representing the Rasch model. The present formulation is more convenient here. As is true for all IRT models, Equation 7 implies that person ability and item difficulty are scaled in the same metric on a common continuum. It also shows that ability and difficulty combine additively to affect performance; that is, the model postulates a compensatory relationship between ability and difficulty.

The difference between the Rasch model and the LLTM is that the latter provides a decomposition of item difficulties of the Rasch model. Data that are "external" to the encounter of the person and the item are required to identify the parameters. Specifically, items must be scored on k complexity factors that are hypothesized to determine item difficulty.

A simple linear model of the complexity factors, Q_{ik}, multiplied by their difficulty, η_k, is assumed to explain item difficulty as follows:

$$\beta_i^* = \sum_k \eta_k Q_{ik} + d \qquad [8]$$

where:

Q_{ik} = the score of item i on factor k or the complexity of factor k in item i

η_k = the weight of factor k in difficulty

d = a normalization constant

The factors that are scored, Q_{ik}, may have either positive or negative weights, η_k, in item difficulty.

This mathematical model for task processes is combined with the Rasch model as formulated in Equation 7. Thus, the full LLTM is written as follows:

$$P(X_{ij}=1/\theta_j,\ \beta_i) = \frac{\exp[\theta_j - (\Sigma_k\ \eta_k\ Q_{ik} + d)]}{1 + \exp[\theta_j - (\Sigma_k\ \eta_k\ Q_{ik} + d)]} \qquad [9]$$

As can be clearly seen in Equation 9, the LLTM combines a mathematical model of component processes or task features (complexity factors) and a psychometric model of individual differences at the item level. (Answering an item correct depends in part on the ability of the person encountering the item.) So, the LLTM links item difficulty to external or theoretical variables and yet retains the properties of the Rasch model as a psychometric model.

A major advantage of the LLTM resides in its capability of testing hypotheses about models of task performance or difficulty. Since it attempts to explain task or item difficulty from a small set of parameters, the LLTM is an explanatory model. The explanatory factors in the LLTM may be either scored variables derived from some theory, or empirical ratings or classifications of the item stimuli on variables that are postulated to influence performance by the target population. The LLTM is a multicomponent model in the sense that multiple stimulus or component factors are hypothesized to underlie task or item difficulty. However, it should be noted that the LLTM is a unidimensional model, as it contains only one ability parameter for a person.

Since the LLTM is an extension of the Rasch model, the assumptions of the Rasch models are also assumptions of the LLTM. Even the assumption of unidimensionality is preserved in the LLTM, which is not the case for the MLTM.

With regard to estimation, G. H. Fischer (1973) has derived conditional maximum likelihood estimators for the item parameters of the LLTM, n_k. But, although conditional maximum likelihood estimators are statistically superior to unconditional estimators for several reasons (see G. H. Fischer, 1981), they are impractical for large sets of items. Hence, the conditional maximum likelihood estimation procedure is useful for the estimation of LLTM parameters when the data set is not large (not over 80 items (Fischer & Forman, 1982)). G. H. Fischer and Forman (1972) have implemented the conditional maximum likelihood estimation method in a computer program. Whitely and Nieh (1981) have adopted the same procedure in their LINLOG computer program.

Normal applications of LINLOG yield both person- and item-parameter estimates for both the LLTM and the Rasch model. The adequacy of the complexity-factor model for item difficulty may be evaluated in three ways: (i) by a log-likelihood χ^2 for differences in goodness-of-fit between the LLTM and the Rasch model. The χ^2 value is -2 times the likelihood difference between the two models; (ii) by correlations of item difficulties $\beta*$ and β obtained, respectively, under the restricted LLTM and the Rasch model; and (iii) by testing the significance of each complexity factor, since standard errors and error correlations are also obtained in a LINLOG analysis.

In LINLOG, the conditional maximum likelihood estimates for the item parameters are computed by the gradient search method described by G. H. Fischer and Forman (1972). Person parameters are estimated by the Newton-Raphson method, after the item parameters habe been obtained.

Research applications with the LLTM

Several investigations have been done using the LLTM as a modeling tool. An early example of the application of the LLTM is G. H. Fischer's (1973) study of the procedures required to solve problems in the differentiation of explicit functions, as taught in secondary schools in Austria. Eight basic cognitive operations that the examinees must apply to solve the problems were derived from an analysis of the structure of the calculus test. These operations were hypothesized to account for the relative difficulties of the problems. Fischer found an agreement between the Rasch item-difficulty parameters and a linear combination of the occurrence and difficulty of the hypothesized operations. In other words, the LLTM parameters could explain the Rasch item difficulty parameters, as hypothesized by the investigator.

Another example of the application of the LLTM is Whitely and Schneider's (1981) study of models of the course of information processing for geometric analogies. These researchers tested three alternative accounts of processing on geometric analogies.

Model I in their study represented Mulholland, Pellegrino, and Glaser's (1980) information-structure variables. Mulholland et al. postulated that two processing variables are crucial in the solution of geometric analogies, namely, encoding complexity and transformation complexity. The former is determined by the number of elements in the first stimulus in an analogy; the latter depends on the number of transformations required to transform the first stimulus to the second stimulus. (See Whitely & Schneider, 1981; p. 384, for an example of a geometric analogy.) Hence, each item can be scored on the two complexity factors, encoding and transformation. In the study, Model II made a distinction between figure distortion transformations (e. g., a change in shape) and displacement transformations (e. g., rotation). Finally, Model III considered several types of transformations.

Likelihood ratio chi-square tests were used to compare the goodness-of-fit of the models. Model II was found to fit the data better than Model I and Model III was better than Model II, although the difference between the latter two was trivial. For parsimony, Model II was selected as the best of the three models.

Whitely and Schneider (1981) also evaluated the ability of Model III to account for Rasch item-difficulty parameters. Note that the Rasch model can be considered as a saturated model, since it can be used to reproduce the item-difficulty data. Model III, the most complex information-processing model estimated for the data, was significantly different from the Rasch model. That conclusion was reached because the obtained LLTM parameters could not account for the item difficulties. This implies that other types of complexity factors should be considered for geometric analogies.

More recently, another investigator, Medina-Diaz (1993), used the LLTM to define the cognitive structure of an algebra test involving one-variable linear equations. Eight production rules representing the mathematical procedures required to solve the problems described the cognitive structure of the test. Another technique, quadratic assignment (QA, Hubert & Schultz, 1976) was employed to validate the hypothesized structure, that is, to ascertain that examinees solved the problems using the specified production rules. This research shows that there are benefits in applying both the LLTM and QA in test analysis.

G. H. Fischer (1978) has provided several examples of European studies with the LLTM.

In the present study, the LLTM was used as a modeling tool to investigate the structure of simple addition and subtraction word problems as hypothesized by current models of children's performance on those problems.

E. Summary and rationale for the present study

To summarize, the factors underlying the relative difficulty of addition and subtraction word problems have been of paramount importance, over the years, to researchers and educators interested in the development of children's problem-solving skills in arithmetic. As was evident in the literature review, to many of the scholars currently involved in determining these factors, the question of how children interpret and represent arithmetic word problems deserves particular attention. This makes sense, since children's ability to understand arithmetic problem texts antecedes their ability to apply problem-solving procedures. This is why research perspectives based on knowledge structures have taken center stage in investigations of children's arithmetic. In this respect, the distinction between conceptual knowledge and procedural knowledge has been emphasized by researchers.

The question of the role of conceptual knowledge and linguistic knowledge in the development of children's skills in solving addition and subtraction word problems has been the focus of many investigations. Various approaches have been used in these

investigations, including retelling data (e.g., Cummins et al., 1988; Stern, 1993; Verschaffel, 1994), problem rewording (e.g., Cummins, 1991; Davis-Dorsey et al., 1991), computer simulation (e.g., Riley & Greeno, 1988), training studies (e.g., F. E. Fischer, 1990; Lewis, 1989), and eye-movement studies (e.g., De Corte, Verschaffel, & Pauwells, 1990; Hegarty, Mayer, & Green, 1992; Verschaffel, De Corte, & Pauwells, 1992).

From previous studies, it seems clear that some characteristics of the problems themselves (e.g., semantic structure, linguistic complexity) and children's conceptual and procedural knowledge influence children's solution success on the problems. But, what are these characteristics of the problems?

The position taken here was that a modeling approach based on the LLTM could be useful in that it would make it possible to use the structural features of the problems as a basis for explaining children's performance levels on the problems.

III. METHOD

A. Subjects

Forty children from each of Grades 1, 2, and 3, with an equal number (i.e., 20) of girls and boys at each grade level, served as participants in the study. Mean ages were 6 years 5 months, 7 years 6 months, and 8 years 5 months, respectively, for first graders, second graders, and third graders. The children were attending a large, predominantly white middle-class elementary school located in an outlying district of a large Western Canadian city. For all children participating in the study, English was the native language.

The study was conducted during the first half of the school year, in November and December. Written parental consent was secured for each child's participation in the study.

B. Tasks and materials

Thirty-two arithmetic word problems were used as test problems in the study. The problems were constructed to represent each of the four major categories of problems that researchers have used to chart the domain of simple addition and subtraction word problems; namely, change, combine, compare, and equalize problems. In each problem, three different sets were mentioned: a start set, a change or transfer set, and a result set. But, the number of objects in only two of the sets was stated in each problem.

In change problems, a start set undergoes a transfer-in or transfer-out of items; then, a question is asked about the cardinality of one of the three sets. In combine problems, the cardinality of a subset or superset must be computed given information about the other two sets. In compare problems, a question is asked about the cardinality of one set, which must be computed by comparing the information given about the relative sizes of the other sets. Equalize problems are thought of as a hybrid category, as they have features of both change and compare problems. Specifically, there is the same sort of action as in change problems, but that action is based on the comparison of two disjoint sets. Hence, all the problems were one-step problems that could be solved either by addition or by subtraction by using two of the numbers stated in the problems. A more detailed description of each of these categories of problems can be found in the literature review.

Each problem was stated in three or four sentences. In problems consisting of three sentences, the first two sentences contained information about the number of objects in two of the sets. For problems consisting of four sentences the information about the cardinality of two of the sets was contained in the first three sentences. For these problems, one sentence among the first three sentences did not refer to a specific number as the cardinality of one of the sets; rather, the sentence always contained the word "some", which was a reference to the unspecified number in the set whose cardinality had to be computed. Invariably, the last sentence in each problem was a question about the cardinality of one of the three sets mentioned in the problem.

The problems were constructed from two problem frames. A problem frame involves the same kinds of objects, related to the same locations or individuals. In the present study, one frame involved marbles, ownership, Connie, and Eric, as exemplified by the following problem:

> *Eric had 7 marbles. Then Connie gave him 3 more marbles. How many marbles does Eric have now?*

Twenty-four problems were prepared from the first frame. These problems consisted of six instances within each of the four major categories of arithmetic word problems. The other problem frame involved cows, ownership, Butcher Joe, and Farmer Pete, as illustrated by the following problem:

> *Butcher Joe had some cows. Then Farmer Pete sold him 3 cows. Now Butcher Joe has 7 cows. How many cows did Butcher Joe have in the beginning?*

Eight problems were constructed from the second frame. These problems consisted of two instances within each of the four major problem categories. For the second frame, only two instances were used within each problem type for two reasons. The first reason was to keep testing time to a reasonable level for children of these age groups. The second reason was to present more of the most difficult problems, since these problems were of most interest in the present study.

The use of two problem frames to construct the test problems for the present study was a departure from studies in which a different frame was employed for each problem (e.g., Riley & Greeno, 1988), and those in which the same frame was used for all problems (e.g., Cummins et al., 1988). The present investigator thought that using a different frame for each problem could be distracting to subjects or could increase the memory load required by the tasks, as subjects would have to attend to different "actors" and "objects" in order to comprehend the problems. Note that for successful performance on the test problems, children only need to attend to the relationships among the sets and to remember the numbers stated in the problems to describe the absolute or relative sizes of the sets. It was also believed that the use of only one frame for all problems could bore subjects as they would have to hear the "same story" over and over again. The 32 problems are presented in Tables 2 and 3.

For any problem, the two numbers used to refer to the sets whose cardinalities were given, included only the numbers 2 through 10. Small numbers were used in the problems in keeping with previous studies (e.g., Cummins, 1991; Cummins et al., 1988; Riley & Greeno, 1988; Riley et al., 1983) that have investigated children's performance on arithmetic word problems. In these studies, the practice has been to deemphasize computational complexity resulting from the use of large numbers. Additionally, in the present study, the numbers were selected such that, for any problem, the correct answer was always 10 or less, and always different from the two numbers stated in the problem. This constraint was motivated by young children's widespread use of what has been referred to as a "given number strategy" (e.g., Cummins et al., 1988; p. 417), as frequently reported in the literature (e.g., De Corte et al., 1985; Riley et al., 1983). It was important here, as in previous studies, that the use of that strategy by children produce an error: Had a number given in a problem been the correct answer, it would have been difficult to interpret a correct answer to that problem given by a child.

Table 2. Test Problems Used in the Study

Frame 1

Change Problems

1. Eric had 7 marbles. Then Connie gave him 3 more marbles. How many marbles does Eric have now?
2. Eric had 9 marbles. Then he gave 4 marbles to Connie. How many marbles does Eric have now?
3. Eric had 3 marbles. Then Connie gave him some marbles. Now Eric has 10 marbles. How many marbles did Connie give him?
4. Eric had 8 marbles. Then he gave some marbles to Connie. Now Eric has 5 marbles. How many marbles did he give to Connie?
5. Eric had some marbles. Then Connie gave him 3 marbles. Now Eric has 9 marbles. How many marbles did Eric have in the beginning?
6. Eric had some marbles. Then he gave 4 marbles to Connie. Now Eric has 6 marbles. How many marbles did Eric have in the beginning?

Combine Problems

1. Eric has 6 marbles. Connie has 4 marbles. How many marbles do they have altogether?
2. Eric and Connie have some marbles. Eric has 6 marbles. Connie has 4 marbles. How many marbles do they have altogether?
3. Eric has 2 marbles. Connie has some marbles. Eric and Connie have 10 marbles altogether. How many marbles does connie have?
4. Eric has some marbles. Connie has 5 marbles. Eric and Connie have 9 marbles altogether. How many marbles does Eric have?
5. Eric and Connie have 10 marbles altogether. Eric has 3 marbles. How many marbles does Connie have?
6. Eric and Connie have 10 marbles altogether. Eric has some marbles. Connie has 7 marbles. How many marbles does Eric have?

Table 2. (Continued)

Compare Problems

1. Eric has 3 marbles. Connie has 8 marbles. How many marbles does Connie have more than Eric?
2. Eric has 10 marbles. Connie has 3 marbles. How many marbles does Connie have less than Eric?
3. Eric has 4 marbles. Connie has 5 marbles more than Eric. How many marbles does Connie have?
4. Eric has 8 marbles. Connie has 3 marbles less than Eric. How many marbles does Connie have?
5. Eric has 7 marbles. He has 3 marbles more than Connie. How many marbles does Connie have?
6. Eric has 3 marbles. He has 7 marbles less than Connie. How many marbles does Connie have?

Equalize Problems

1. Eric has 9 marbles. If Connie gets 3 more marbles she will have the same number of marbles as Eric. How many marbles does Connie have?
2. Eric has 10 marbles. If Eric gives 3 marbles away he will have the same number of marbles as Connie. How many marbles does Connie have?
3. Eric has 8 marbles. Connie has 3 marbles. How many more marbles does Connie need to get to have as many marbles as Eric?
4. Eric has 9 marbles. Connie has 5 marbles. How many marbles does Eric need to give away to have as many marbles as Connie?
5. Eric has 4 marbles. If he gets 5 more marbles, he will have the same number of marbles as Connie. How many marbles does Connie have?
6. Eric has 6 marbles. If Connie loses 2 marbles, she will have the same number of marbles as Eric. How many marbles does Connie have?

Table 3. Test Problems Used in the Study

Frame 2

Change Problems

5. Butcher Joe had some cows. Then Farmer Pete sold him 3 cows. Now Butcher Joe has 7 cows. How many cows did Butcher Joe have in the beginning?
6. Butcher Joe had some cows. Then he sent 2 cows to Farmer Pete. Now Butcher Joe has 5 cows. How many cows did Butcher Joe have in the beginning?

Combine Problems

5. Butcher Joe and Farmer Pete have 9 cows altogether. Butcher Joe has 4 cows. How many cows does Farmer Pete have?
6. Butcher Joe and Farmer Pete have 9 cows altogether. Butcher Joe has some cows. Farmer Pete has 5 cows. How many cows does Butcher Joe have?

Compare Problems

5. Butcher Joe has 8 cows. He has 3 cows more than Farmer Pete. How many cows does Farmer Pete have?
6. Butcher Joe has 6 cows. He has 2 cows less than Farmer Pete. How many cows does Farmer Pete have?

Equalize Problems

5. Butcher Joe has 3 cows. If he buys 4 more cows, he will have the same number of cows as Farmer Pete. How many cows does Farmer Pete have?
6. Butcher Joe has 4 cows. If Farmer Pete sells 3 cows, he will have the same number of cows as Butcher Joe. How many cows does Farmer Pete have?

In addition to the test problems, two practice problems were prepared, using Farmer John and Butcher Tom as protagonists and cows as objects. The first practice problem was the following Change 2 problem :

> *Farmer John has 5 cows. Then he sold 2 cows to Butcher Tom. How many cows does Farmer John have now?*

The second practice problem was the following Compare 4 problem:

> *Farmer John has 5 cows. Butcher Tom has 2 cows less than Farmer John. How many cows does Butcher Tom have?*

There were 32 white cards, 30 centimeters long and 5 centimeters wide, on which the response alternatives were printed. More specifically, on each card five numbers were printed, one of which was the correct answer to a given problem. The numbers were equally spaced on each card. Furthermore, the position of the correct answer was chosen so that the correct answer appeared, across all problems, approximately an equal number of times in each position.

Finally, marbles and plastic cows were available to children to work with when solving the problems.

C. Procedure

Children were tested individually in a quiet room at their school during school hours. Following the procedure used in previous studies, all problems were presented orally: They were read by the experimenter to the child, who was required to solve them without benefit of paper and pencil. Problem presentation followed the same order for each child. Twelve problems from the first frame were presented first, followed by four problems from the second frame; then the remaining 12 problems from the first frame, and the other four problems from the second frame, were administered.

Each child participating in the study was brought from his or her classroom to the testing room by the experimenter to whom the child was assigned. Then the experimenter and the child sat on the same side of a small desk on which the testing materials were placed. After a short informal conversation during which she established rapport with the child, the experimenter proceeded with the testing session.

Each session began with instructions, followed by the two practice problems. Children were assisted in solving the practice problems, if necessary. Once it was clear that the child understood the procedure, the experimental session was begun. No help by the experimenters was available to children when children were solving the test problems. Before the practice problems were read to them, children were told that they could use the marbles and the plastic cows to solve the problems. However, not a single child used the objects in solving all the problems. In fact, many children completed all the problems without using any marbles or plastic cows. Children were also allowed to solve the problems at their own pace. Specifically, they were told that there was no time limit set for solving each problem or for completing all the problems. Children took an average of 15 mn to complete the test problems. No children were found to be either exceedingly fast or slow in solving the problems. Most problem statements were repeated twice by the experimenters; a few problems were solved after only one reading; an even smaller number of problems needed to be read more than twice before children could provide answers to them.

Two white female experimenters administered the tasks. Children were randomly assigned to the experimenters with the constraint that the two experimenters test an equal number of children of the same sex at each of Grades 1, 2, and 3. The two experimenters were trained by the principal investigator during a 4-hour session consisting of two parts. In the first part of the session, the experimenters were introduced to the test materials; then they were asked to familiarize themselves with the instructions and the test problems. During this process, the experimenters were encouraged to ask any questions they might have on the instructions and the problems by the principal investigator. Finally, the principal investigator emphasized to the experimenters the importance of keeping the procedure uniform across children.

During the second part of the session, each experimenter administered the tasks to three children (one child from each of Grades 1, 2, and 3), in the presence of the other experimenter and the principal investigator. At the conclusion of each session with a child, the two experimenters and the principal investigator critiqued the administration of the tasks. The practice sessions, in addition to providing guided practice to the experimenters, were an attempt to keep all aspects of the administration of the tasks uniform across children. It should be noted that the data for the children who served in the practice sessions were not used in any analysis.

A key feature of the instructions worth mentioning was that children were discouraged against guessing and impulsive responding. Rather, they were encouraged to "think through" each problem and respond by the best possible answer.

Children were awarded one point for each test problem solved correctly and zero point for each problem solved incorrectly.

Some data were collected also from teachers whose pupils served as participants in the study. They were asked to rate their pupils on mathematical competence on a 5-point scale, on which a larger number indicated more competence.

D. Data-analysis strategies and presentation of results

The statistical analysis of the data was composed of two parts: a preliminary analysis and a main analysis. The preliminary analysis was designed to provide classical item statistics that were used to evaluate the psychometric adequacy of the measuring instrument. Some of these indices (e.g., the point-biserial correlation coefficients) were also used in the assessment of the IRT model assumptions.

The main analysis was carried out in three steps. The objective of the first step in the analysis was to make it possible to compare and contrast the pattern of findings in the present study with the patterns of findings in other studies of children's performance on arithmetic word problems. This step consisted of a descriptive level and an inferential level.

At the descriptive level, children's performance on the problems was presented in detail using relevant summary statistics, with an emphasis on performance level as a function of the following variables: grade, sex, category of problem (e.g., change problems vs. compare problems), and presumed difficulty of problems (e.g., Compare 1 problem vs. Compare 5 problem). At the inferential level of the first step in the data analysis, statistical tests of significance were conducted to ascertain whether or not some of these variables reliably explained performance differences among subject groups and among different problems.

The second step in the analysis of data dealt with the assessment of model-data fit. In IRT applications, the importance of model-data fit cannot be over-emphasized, as the success of any application is predicated on the appropriateness of the model for the data under consideration. As noted earlier, many methods have been proposed for assessing goodness-of-fit in IRT modeling (see, for example, Hambleton, 1989). The safest strategy for an investigator is to use several of these methods, preferably from different categories (i.e., model assumptions, expected model features, and model predictions) when assessing goodness-of-fit. In the present study, several methods related to model assumptions and model predictions were employed.

Since the modeling technology used here was the LLTM, which is a modified Rasch model, the following assumptions were checked: unidimensionality, equal discrimination indices, absence of guessing or minimal guessing, and non-speeded test administration.

Unidimensionality was assessed by several procedures. The first procedure was linear factor analysis (LFA), using both phi correlations and tetrachoric correlations, and two factoring methods, principal-components analysis (PCA) and principal-axes factoring (PAF) .

LFA remains the most commonly used approach for assessing test dimensionality. However, there are some fundamental problems in applying LFA to binary (i.e., 0-1) data. For instance, LFA assumes that the relationship between test scores (the observed variables) and the underlying factors or components (the latent variables) is linear and that the variables are continuous, although it can be shown that this relationship is nonlinear. Hence, applying LFA to binary data amounts to approximating a nonlinear relationship by one that is linear.

LFA was employed in the present research to allow comparisons with previous studies and to provide baseline dimensionality results against which to compare the results of other dimensionality-assessment procedures used here.

In an attempt to determine the number of dominant dimensions underlying the data, as revealed by the LFA, several approaches were taken: Kaiser's criterion (eigenvalue greater than 1), the scree plot, amount of variance accounted for, parallel analysis, and goodness-of-fit statistics based on descriptive measures of the residual correlations. Only parallel analysis needs to be described here, as the other approaches are well known, as they are frequently used in factor analyses.

Parallel analysis (e.g., Cota, Longman, Holden, & Fekken, 1993a; Hays, 1987; Horn, 1965; Humphreys & Montanelli, 1975; Zwick & Velicer, 1986) provides an upper bound for the number of factors to retain in exploratory factor analysis. The method compares eigenvalues from the interitem correlation matrix of the real data on hand, and "criterion" eigenvalues from an interitem correlation matrix of random normal deviates in data based on the same sample size and the same number of variables as the real data.

Horn (1965) suggested that criterion eigenvalues be obtained by generating a few sets of random data and computing mean eigenvalues for each ordinal position. Currently, computer programs (e.g., Hays, 1987; Lautenschlager, 1989; Longman, Cota, Holden, & Fekken, 1989a) are available for direct computation or estimation of mean criterion eigenvalues. However, there is some evidence (e.g., Zwick & Velicer, 1986) that parallel analysis shows a slight bias toward the retention of too many components or factors, especially when mean criterion eigenvalues are used. This finding prompted Longman et al. (1989b) to suggest using 95th percentile eigenvalues, which logically should lead to slightly more conservative results than the mean eigenvalue approach. Cota et al. (1993b) provided an empirical example of the interpolation of 95th percentile eigenvalues from random data.

In parallel analysis, factors of the matrix of real data that have eigenvalues greater than those of the comparison matrix of random data are retained. In other words, the number of eigenvalues of the real data greater than the largest eigenvalue of the random data is taken as the number of significant factors underlying the data under consideration.

In a comparison of five criteria for determining the number of components to retain, Zwick and Velicer (1986) found that parallel analysis was one of the two best performing rules across seven systematically varied data conditions. The use of parallel analysis in dimensionality assessment has been suggested by various investigators (e.g., Hambleton, 1989; Silverstein, 1987, 1990; Zwick & Velicer, 1986). In the present research, parallel analysis was performed using Hays's (1987) PARALLEL computer program.

The second procedure used for assessing unidimensionality was nonlinear factor analysis (NLFA), using two different computer implementations, NOHARM (Fraser, 1981; Fraser & McDonald, 1988) and NOFA (Etezadi-Amoli & McDonald, 1983). Note that NLFA "fits nicely" with IRT, that is, NLFA is conceptually consistent with IRT, because IRT models are nonlinear in nature (McDonald, 1982).

Following the recommendations of Hattie (1985) and McDonald (1981, 1982), a one-factor NLFA model was fitted to the interitem correlation matrices (phis and tetrachorics) and the residual correlations were examined. Several NLFA models of high dimensionality (i.e., more than one factor) also were fitted to the data. This allowed to examine the resulting residual correlations. Hambleton and Rovinelli (1986) have successfully applied these procedures in assessing the dimensionality of a set of test items. Fitting several NLFA models of high dimensionality also made possible to compute De Champlain and Gessaroli's (1991) incremental fit index (IFI) based on the sum of squares of the residual covariances obtained in the NOHARM analyses. These investigators define the IFI in the context of assessing the dimensionality of a set of test items as follows:

$$IFI_k = \frac{SS_{res}\ (k\text{-}factor) - SS_{res}\ [(k+1)\text{-}factor]}{SS_{res}\ (k\text{-}factor)} \qquad [10]$$

As can be seen from Equation 10, the IFI computes the proportion of the sum of squares of the residual covariances from the k-factor solution that is accounted for by the (k + 1)th factor. For example, if k = 1, then the value of IFI is the proportion of the sum of squares of the residual covariances from the one-factor solution that is due to the second factor. The advantages of the IFI have been noted by its proponents and are twofold. First, dimensionality assessment is carried out using a model (NLFA) on which IRT is based. Second, the index of model misfit is directly related to the mathematical function minimized in the estimation procedure.

The third procedure used to assess unidimensionality was Stout's (1987, 1990) statistical significance test of essential unidimensionality (i.e., existence of one dominant dimension in the data), as implemented in the DIMTEST computer program. The test hypotheses are stated as:

$$H_0 : d_E = 1 \text{ vs. } H_1 : d_E > 1,$$

where d_E denotes the essential dimensionality of the item domain. The test statistic is the unidimensionality statistic T given by:

$$T = \frac{(T_1 - T_2)}{\sqrt{(2)}} \qquad [11]$$

For the meaning of T_i in Equation 11 and the derivation of the T statistic, see Stout (1987, 1990), Nandakumar (1991, 1993), and Nandakumar and Stout (1993).

The computed T value is referred to the upper tail of the unit normal distribution. Consistent with the above hypotheses, the p-values of essentially unidimensional tests are expected to be large whereas those of multidimensional tests are expected to be less than or equal to the specified level of significance. Stout's T statistic has been shown to discriminate well between essentially unidimensional and multidimensional sets of test scores, for both simulated (Nandakumar, 1994; Nandakumar & Stout, 1993; Stout, 1987) and real data (Nandakumar, 1993, 1994).

For the assessment of the assumption of equal discrimination indices, the point biserial correlations between items and total test and between items and subtests were evaluated for homogeneity.

For the assessment of the assumption of minimal guessing, an analysis of the performance of low-ability students (based on their total score on the test) was done to see whether performance levels are close to zero, which would lend support to the viability of the assumption. Other considerations were important here, including item format, the instructions given to the participants, and the observations by the data collectors.

The assumption of nonspeeded test administration was assessed by referring to relevant data-collection procedures and the observations by the two experimenters.

Concerning model predictions, the assessment focused on the investigation of residuals and standardized residuals after a relevant model was fitted to the data.

The third step in the data analysis dealt with the LLTM analysis. The goal of this analysis was to evaluate the knowledge requirements (i.e., logico-mathematical knowledge and linguistic knowledge) for solving arithmetic word problems suggested in the literature. The 32 word problems used in the present study were dichomously scored on three aspects of logico-mathematical knowledge and three aspects of linguistic knowledge requirements.

The three aspects of logico-mathematical knowledge were part-whole schema, double-role vs. single-role counters, and re-representation. The three aspects of linguistic knowledge requirement were comparative terms, action cues vs. no action cues, and consistent language vs. conflicting language. These characterizations of arithmetic word problems were described and explained in the literature review. It should be noted that the distinction between consistent and conflicting language was not limited to compare problems as in Briars and Larkin (1988) or in Lewis and Mayer (1987); rather it was applied to all the problem types used in the present study.

A problem was scored 1 when a knowledge aspect was required for solving the problem or when a structural or processing feature making the problem more difficult was present. It was scored 0 when a knowledge aspect was not required for solving the problem or when a feature making the problem more difficult was absent.

These dichotomous item scores made up a 32 X 6 weight matrix Q required for the LLTM analysis. Table 4 presents the vector of weights corresponding to each test problem used in the study. Using the LINLOG program (Whitely and Nieh, 1981), conditional maximum likelihood estimates of the η_k and β_i* parameters were obtained, and significance tests of the parameters corresponding to the structural features of the problems were carried out. These significance tests evaluated the importance of the structural features in determining problem difficulty.

Table 4. Item Sequence, Item Code and the Q Matrix

		Structural And Processing Complexity (Q Matrix)					
Item Sequence	Item Code	Part-Whole	Double-Role Counters	Re-Representation	Comparative Terms	Action Cues	Language Constancy
1	CH1	0	0	0	0	1	0
2	CH5	1	0	0	0	1	1
3	CH3	0	1	0	0	1	1
4	CH6	1	0	1	0	1	1
5	CH4	0	1	0	0	1	0
6	CH2	0	0	1	0	1	0
7	CP5	1	0	0	1	0	1
8	CP1	0	0	0	1	0	0
9	CP3	0	0	0	1	0	0
10	CP2	0	0	0	1	0	0
11	CP6	1	0	0	1	0	1
12	CP4	0	0	0	1	0	0
13	EQ6	0	0	0	0	1	1
14	CB5	1	1	0	0	1	0
15	EQ5	0	0	0	0	1	0
16	CB6	1	0	0	0	1	0
17	CB1	0	0	0	0	1	0
18	CB5	1	1	0	0	1	0
19	CB3	1	1	0	0	1	0
20	CB6	1	0	0	0	1	0
21	CB4	1	0	0	0	1	0
22	CB2	0	0	0	0	1	0
23	EQ1	0	0	0	0	1	1
24	EQ5	0	0	0	0	1	0
25	EQ3	0	0	0	0	1	1
26	EQ6	0	0	0	0	1	1
27	EQ4	0	0	0	0	1	0
28	EQ2	0	0	0	0	1	0
29	CH5	1	0	1	0	1	1
30	CP6	1	0	0	1	0	0
31	CH6	1	0	1	0	1	1
32	CP5	1	0	0	1	1	0

IV. RESULTS

This chapter provides a detailed presentation of the results of the statistical analyses described in the preceding part of the thesis. This presentation involves two parts, the first dealing with the preliminary analysis and the second focusing on the main analysis.

A. Preliminary analysis: sizing up the measuring instrument

This section reports the results of the analysis carried out to obtain classical item statistics. Tables 5, 6, 7, and 8 contain these statistics, respectively, for change, combine, compare, and equalize problems. Each table gives the following information: order of presentation of the item (sequence number), type of item (item code), difficulty and discrimination indexes, point-biserial correlation coefficients of the item with the total test and with the subtest to which the item belongs, the proportion of the total, high, and low groups selecting each alternative, the discrimination index for each alternative, the average total test score and the average subtest score for examinees selecting a given option.

The test used in the present study was not developed using any elegant psychometric procedures. However, an inspection of the item statistics obtained in the preliminary analysis shows that the numbers are within acceptable limits or are reasonable for what they indicate. For instance, item difficulties vary from a high of .83 (for Change 1) to a low of .39 (for Compare 6), the discrimination indexes are all higher with the corresponding point-biserial correlation coefficients with the total test; all point-biserial correlations with the subtest are higher with the corresponding point-biserial correlations with the total test; all discrimination indexes for the distractors are negative numbers, that is, all distractors show negative discrimination, which means that they were chosen more often by members of the low group than by those of the high group; all the average total test scores and average subtest scores for examinees choosing the correct option are higher than the corresponding averages for examinees choosing any other alternative. All these patterns are expected if the test is adequate.

Some descriptive information about the classical psychometric indexes of the test items is presented in Table 9.

The analysis considered the four major categories of arithmetic word problems as component subtests. Based on that decomposition of the test, subtest statistics were computed and are shown in Table 10. There do not seem to be any marked differences among change, compare, and equalize problems; however, all these problem types seem to be different from compare problems. A subtest correlation matrix was also obtained and can be seen in Table 11.

For the whole instrument, Cronbach's alpha for internal consistency reliability was .92, and the standard error was 2.312. Thus, the instrument had good internal consistency reliability.

Table 5. Classical Item Analysis Results for Change Problems

Item statistics						Alternative statistics						
Sequence number	Item code	Diff. index	Discr. index	Pt. bis. (total test)	Pt. bis. (subtest)	Alt.	Prop. (total)	select. (high)	alt. (low)	Discr. index	Means (total test)	Means (subtest)
1	CH1	.83	.34	.38	.44	1	.05	.00	.16	-.16	9.67	2.00
						2*	.83	1.00	.66	.34	19.49	5.31
						3	.07	.00	.09	-.09	12.56	3.11
						4	.03	.00	.06	-.06	11.00	3.75
						5	.02	.00	.03	.01	5.00	1.00
6	CH2	.78	.50	.49	.65	1*	.78	1.00	.50	.50	20.20	5.63
						2	.08	.00	.13	-.13	12.80	2.60
						3	.08	.00	.22	-.22	9.20	2.00
						4	.02	.00	.06	-.06	9.67	2.00
						5	.02	.00	.09	-.09	7.00	2.00
3	CH3	.59	.72	.58	.66	1	.07	.00	.07	-.07	13.78	3.33
						2	.03	.00	.03	-.03	16.25	4.25
						3	.08	.06	.09	-.03	17.10	4.40
						4*	.59	.94	.22	.72	22.04	6.07
						5	.22	.00	.59	-.59	9.38	2.50
5	CH4	.67	.56	.44	.57	1	.06	.03	.19	-.16	9.43	2.00
						2	.09	.10	.16	-.06	16.82	3.91
						3*	.67	.87	.31	.56	20.66	5.75
						4	.02	.00	.00	.00	16.50	4.50
						5	.17	.00	.34	-.34	11.60	3.05
2	CH5	.50	.65	.49	.48	1	.09	.06	.12	-.06	15.55	4.09
						2	.32	.06	.50	-.44	14.18	3.84
						3	.07	.00	.13	-.13	11.63	3.63
						4	.02	.00	.03	-.03	12.00	3.67
						5*	.50	.87	.22	.65	22.17	5.93
29	CH5	.45	.72	.55	.65	1*	.45	.81	.09	.72	23.11	6.43
						2	.09	.00	.13	-.13	13.45	4.27
						3	.17	.00	.38	-.38	11.20	2.55
						4	.17	.16	.19	-.03	17.19	4.67
						5	.12	.03	.22	-.19	13.43	3.14
4	CH6	.51	.59	.46	.58	1	.03	.03	.03	.00	15.50	4.25
						2	.08	.00	.19	-.19	11.60	2.90
						3	.16	.03	.25	-.22	13.32	3.37
						4*	.51	.81	.22	.59	21.84	6.11
						5	.22	.13	.31	-.18	15.62	4.00
31	CH6	.56	.71	.53	.55	1	.11	.00	.19	-.19	14.54	3.92
						2	.07	.00	.13	-.13	13.67	3.11
						3	.14	.03	.25	-.22	12.41	3.29
						4*	.56	.87	.16	.71	21.96	5.94
						5	.12	.10	.28	-.18	12.50	3.86

- Underlined item codes indicate Frame 2-problems.
- For each item, the alternative with the asterisk is the keyed option.

Table 6. Classical Item Analysis Results for Combine Problems

Item statistics						Alternative statistics						
Sequence number	Item code	Diff. index	Discr. index	Pt. bis. (total test)	Pt. bis. (subtest)	Alt.	Prop. (total)	select. (high)	alt. (low)	Discr. index	Means (total test)	Means (subtest)
17	CB1	.87	.22	.28	.33	1*	.87	.97	.75	.22	18.97	5.25
						2	.04	.00	.06	-.06	13.40	2.80
						3	.02	.00	.09	-.09	6.67	2.00
						4	.02	.00	.03	-.03	14.67	4.33
						5	.04	.03	.06	-.03	13.00	2.20
22	CB2	.64	.84	.68	.72	1	.07	.00	.19	-.19	9.88	2.88
						2	.22	.00	.44	-.44	11.54	2.31
						3*	.64	1.00	.16	.84	22.26	6.30
						4	.02	.00	.06	-.06	9.67	3.00
						5	.05	.00	.16	-.16	7.83	2.17
19	CB3	.61	.65	.54	.69	1	.13	.03	.19	-.16	14.56	3.38
						2*	.61	.94	.29	.65	21.70	6.36
						3	.06	.00	.09	-.09	14.29	2.86
						4	.10	.03	.19	-.16	11.83	2.67
						5	.10	.00	.25	-.25	9.17	1.67
21	CB4	.53	.81	.68	.71	1	.14	.00	.22	-.22	13.24	3.65
						2	.06	.00	.13	-.13	10.14	2.57
						3	.14	.00	.31	-.31	10.76	2.18
						4	.13	.06	.22	-.16	12.93	3.20
						5*	.53	.94	.13	.81	23.38	6.64
18	CB5	.68	.62	.57	.70	1	.07	.00	.09	-.09	11.88	2.13
						2*	.68	.94	.31	.63	21.27	6.15
						3	.02	.00	.03	-.03	12.67	2.67
						4	.15	.03	.38	-.35	10.56	2.17
						5	.07	.03	.19	-.16	11.33	2.44
14	CB5	.65	.66	.55	.67	1	.05	.00	.06	-.06	11.67	1.83
						2	.04	.00	.03	-.03	15.60	4.00
						3	.19	.00	.44	-.44	11.30	2.30
						4*	.65	1.00	.34	.66	21.42	6.19
						5	.07	.00	.13	-.13	11.25	2.88
20	CB6	.48	.91	.71	.83	1	.17	.00	.31	-.31	12.62	2.52
						2	.15	.00	.22	-.22	14.50	3.72
						3	.07	.00	.13	-.13	11.00	2.22
						4*	.48	.97	.06	.91	24.17	7.12
						5	.12	.03	.28	-.25	10.14	2.64
16.	CB6	.45	.88	.72	.76	1	.04	.00	.06	-.06	12.20	1.20
						2*	.45	.94	.06	.88	24.65	7.07
						3	.14	.00	.19	-.19	13.41	2.94
						4	.17	.03	.31	-.28	12.76	3.81
						5	.19	.03	.38	-.35	12.22	3.13

- Underlined item codes indicate Frame 2-problems.
- For each item, the alternative with the asterisk is the keyed option.

Table 7. Classical Item Analysis Results for Compare Problems

Item statistics						Alternative statistics						
Sequence number	Item code	Diff. index	Discr. index	Pt. bis. (total test)	Pt. bis (subtest)	Alt.	Prop. (total)	select. (high)	alt. (low)	Discr. index	Means (total test)	Means (subtest)
8	CP1	.57	.81	.61	.72	1	.08	.00	.09	-.09	14.80	1.90
						2	.06	.00	.09	-.09	17.71	2.00
						3*	.57	.97	.16	.81	22.49	5.40
						4	.10	.00	.16	-.16	15.00	2.33
						5	.19	.03	.50	-.47	10.00	0.78
10	CP2	.58	.91	.66	.72	1	.33	.00	.78	-.78	10.80	1.20
						2	.02	.00	.03	-.03	16.00	2.67
						3	.04	.00	.06	-.06	15.00	3.00
						4	.02	.00	.03	-.03	13.50	1.00
						5*	.58	1.00	.09	.91	22.67	5.33
9	CP3	.40	.74	.58	.67	1*	.40	.81	.06	.74	23.98	5.90
						2	.05	.03	.06	-.03	16.83	3.17
						3	.09	.00	.22	-.22	10.73	1.55
						4	.36	.10	.53	-.43	14.09	2.14
						5	.10	.06	.13	-.06	16.08	2.92
12	CP4	.58	.81	.63	.71	1	.07	.06	.13	-.07	14.50	2.38
						2*	.58	.94	.13	.81	22.57	5.35
						3	.19	.00	.41	-.41	11.70	1.00
						4	.07	.00	.19	-.19	9.89	1.44
						5	.09	.00	.16	-.16	12.55	2.00
7	CP5	.43	.78	.62	.75	1	.20	.00	.22	-.22	14.67	2.50
						2	.05	.03	.09	-.06	12.17	2.00
						3	.09	.03	.19	-.16	11.73	1.82
						4*	.43	.87	.09	.78	24.04	6.02
						5	.23	.06	.41	-.35	13.89	1.68
32	CP5	.40	.78	.55	.58	1*	.40	.81	.03	.78	23.69	5.60
						2	.24	.10	.31	-.21	15.21	2.66
						3	.07	.03	.09	-.06	16.78	3.44
						4	.17	.03	.34	-.31	12.20	1.80
						5	.12	.03	.22	-.19	14.00	2.36
11	CP6	.38	.74	.59	.67	1	.28	.16	.44	-.28	14.56	2.50
						2	.04	.00	.06	-.06	14.60	2.00
						3	.05	.03	.06	-.03	16.83	2.67
						4	.25	.06	.44	-.38	13.40	2.17
						5*	.38	.74	.00	.74	24.40	6.00
30	CP6	.39	.62	.42	.59	1	.08	.00	.16	-.16	11.70	0.90
						2	.03	.00	.06	-.06	11.50	2.75
						3	.03	.00	.09	-.09	8.50	0.75
						4*	.39	.68	.06	.62	22.45	5.66
						5	.46	.32	.63	-.31	16.67	2.85

- Underlined item codes indicate Frame 2-problems.
- For each item, the alternative with the asterisk is the keyed option.

Table 8. Classical Item Analysis Results for Equalize Problems

Item statistics						Alternative statistics						
Sequence number	Item code	Diff. index	Discr. index	Pt. bis. (total test)	Pt. bis. (subtest)	Alt.	Prop. (total)	select (high)	alt. (low)	Discr.	Means (total test)	Means (subtest)
23	EQ1	.43	.52	.43	.59	1*	.43	.71	.19	.52	22.20	6.10
						2	.13	.10	.13	-.03	16.60	3.60
						3	.12	.13	.06	-.07	20.14	4.71
						4	.20	.06	.34	-.28	13.75	3.17
						5	.13	.00	.28	-.28	11.00	2.44
28	EQ2	.67	.46	.40	.50	1	.07	.03	.09	-.06	15.88	4.38
						2	.05	.10	.00	.10	23.00	5.50
						3*	.67	.87	.41	.46	20.42	5.35
						4	.07	.00	.19	-.19	10.78	1.56
						5	.14	.00	.31	-.31	10.18	2.12
25	EQ3	.53	.77	.68	.73	1	.22	.00	.44	-.44	10.88	2.77
						2	.11	.00	.28	-.28	9.00	1.85
						3	.08	.03	.03	.00	17.10	3.50
						4	.07	.06	.13	-.07	14.88	3.63
						5*	.52	.90	.13	.77	23.48	6.13
27	EQ4	.73	.53	.51	.63	1*	.73	.97	.44	.53	20.64	5.41
						2	.02	.00	.09	-.09	5.67	0.67
						3	.12	.03	.22	-.19	12.79	2.43
						4	.02	.00	.03	-.03	12.67	3.00
						5	.10	.00	.22	-.22	9.92	2.08
24	EQ5	.60	.53	.43	.55	1	.11	.06	.22	-.16	13.38	3.15
						2	.02	.00	.03	-.03	11.00	2.50
						3	.17	.13	.22	-.09	15.62	3.19
						4*	.60	.81	.28	.53	20.99	5.57
						5	.10	.00	.25	-.25	11.17	2.67
15	EQ5	.69	.46	.42	.54	1	.07	.03	.16	-.13	11.50	2.75
						2	.07	.00	.19	-.19	10.63	2.50
						3	.11	.10	.19	-.09	14.23	2.69
						4	.07	.03	.09	-.06	14.25	3.00
						5*	.69	.84	.38	.46	20.40	5.36
26	EQ6	.48	.55	.48	.57	1	.16	.06	.31	-.25	12.79	2.47
						2*	.48	.74	.19	.55	22.28	5.91
						3	.22	.19	.28	-.09	15.67	3.59
						4	.06	.00	.03	-.03	17.00	4.71
						5	.08	.00	.19	-.19	11.40	3.20
13	EQ6	.43	.78	.58	.59	1	.01	.00	.03	-.03	8.00	3.00
						2	.48	.13	.84	-.71	13.55	3.26
						3	.02	.03	.01	.02	23.00	5.50
						4*	.43	.84	.06	.78	23.58	6.08
						5	.06	.00	.06	-.06	14.71	3.86

- Underlined item codes indicate Frame 2-problems.
- For each item, the alternative with the asterisk is the keyed option.

Table 9. Descriptive Information About Classical Psychometric Indexes of Test Items

Index	Mean	Median	Standard Deviation	Maximum	Minimum
Discrimination	.66	.63	.16	.91	.22
Difficulty	.57	.57	.13	.87	.38
Point-biserial with total test	.54	.55	.11	.72	.28
Point-biserial with subtest	.62	.65	.10	.83	.33

Table 10. Subtest Statistics

Subtest	Mean	Standard Deviation	Highest Score	Lowest Score	Cronbach α	Standard Error of Measurement
Change Problems	4.89	2.16	8	0	.71	1.16
Combine Problems	4.92	2.59	8	0	.84	1.03
Compare Problems	3.72	2.67	8	0	.83	1.10
Equalize Problems	4.55	2.27	8	0	.73	1.18

Table 11. Subtest Correlation Matrix

	Change Problems	Combine Problems	Compare Problems	Equalize Problems
Change Problems	--------			
Combine Problems	.68	--------		
Compare Problems	.67	.64	--------	
Equalize Problems	.63	.60	.67	--------

B. Main analysis

1. Describing children's performance on the problems

The proportions of correct solutions given for each of the 32 test problems are presented in Table 12. As can be clearly seen, there were large differences among problems within problem categories. For instance, the Change 1 problem was solved by 83% of the children whereas the two Change 5 problems were solved by 48% of the children; likewise, 87% of the children could solve the Combine 1 problem, but only 47% could solve the two Combine 6 problems.

Table 12. Proportions of Correct Solutions as a Function of Grade

Item Code	Grade 1	Grade 2	Grade 3	Average
CH1	.80	.80	.90	.83
CH2	.60	.80	.93	.78
CH3	.40	.63	.75	.59
CH4	.58	.62	.78	.66
CH5	.32	.46	.67	.48
CH6	.46	.61	.81	.63
All change problems	.53	.65	.81	.66
CB1	.78	.90	.93	.87
CB2	.53	.55	.83	.64
CB3	.45	.65	.73	.61
CB4	.43	.45	.78	.55
CB5	.49	.69	.84	.67
CB6	.26	.50	.66	.47
All combine problems	.49	.62	.78	.64
CP1	.40	.50	.80	.57
CP2	.30	.63	.83	.58
CP3	.20	.38	.63	.40
CP4	.30	.58	.85	.58
CP5	.19	.43	.64	.42
CP6	.18	.39	.60	.39
All compare problems	.26	.49	.73	.49
EQ1	.32	.33	.63	.43
EQ2	.53	.75	.73	.67
EQ3	.30	.48	.80	.53
EQ4	.70	.58	.93	.73
EQ5	.48	.73	.73	.65
EQ6	.23	.43	.68	.45
All equalize problems	.43	.55	.75	.58
All problems	.43	.58	.77	.59

. For item codes with the digits 5 and 6, the proportions are averages based on Frame-1 and Frame-2 problems.

Large differences were also found among the problems in different categories. For example, some of the change and combine problems (e.g., Change 1 and Combine 1 problems) were solved by most of the children, whereas most of the compare and equalize problems were solved by only a few of the children. The compare problems were the most difficult for the children. To illustrate, the Compare 6 problems were solved by only 18% of the first graders, 39% of the second graders, and 60% of the third graders. Overall, change problems were solved by 66% of the children, combine problems by 64% of the children, equalize problems by 59% of the children, and compare problems by 49% of the children.

In general, the performance levels of the children in the present research were consistent with findings from previous studies of children's performance on arithmetic word problems (e.g., Riley & Greeno, 1988).

An initial analysis of variance using experimenter (Experimenter K vs. Experimenter L) and sex of participants as independent variables was conducted in an attempt to determine whether or not these variables had any effects on children's performance on the problems. No significant effects were found for sex of participants and for experimenter, first using performance on all the problems as the dependent variable; then using performance on problems within the same category as the dependent variable. Consequently, these two variables were collapsed in all the remaining analyses.

Then, a profile analysis was performed on the four problem categories (i. e., change, combine, compare, and equalize problems). The grouping variable was grade. Profile analysis is an application of multivariate analysis of variance (MANOVA) in which several dependent variables are measured on the same scale. In the present study, the four problem categories were considered the dependent variables, and they were measured on the same scale. Hence, profile analysis was appropriate.

The assumption that the variance-covariance matrices for the transformed variables for a particular effect be equal for all levels of the between-subjects factor was examined using the multivariate generalization of Box's M test. The p-value for this test was .08, implying that the assumption was tenable in the present data.

Using Wilks's criterion, the profiles did not deviate significantly from parallelism, $\lambda = .91$; $F(6, 230) = 1.81$, $p > .05$. For the levels test, reliable differences were found among groups, $F(2, 117) = 16.66$, $p < .001$. The effect size estimate η^2 was .22, indicating a moderate association between grade levels and averaged performances on the problems. The final omnibus test was an evaluation of the flatness hypothesis. This hypothesis was rejected by all multivariate criteria, all of which showed essentially the same result. For instance, Wilks' $\lambda = .72$, $F(3, 115) = 14.93$, $p < .001$. The multivariate effect size estimate η^2 was .28.

Since only the levels and the flatness tests were significant, contrasts were performed on the marginal means. Specifically, after the significant levels test, contrasts were done on the marginal values for the grouping variable, grade. Hence, these comparisons were among the overall means (i.e., on the 32 problems) for Grades 1, 2, and 3. Children in Grade 3 had higher means than children in Grades 1 and 2.

Following the significant flatness test, contrasts were performed on the marginal values for the four dependent variables (i.e., the means for change, combine, compare, and equalize problems). The means for compare problems were lower than the means for the other problem categories.

A Pearson correlation coefficient was computed between overall performance on the problems and teachers' ratings of the children's mathematical competence. This correlation coefficient was significant, r (118) = .49, $p < .01$, indicating that the two variables were positively related.

2. Assessing goodness-of-fit of the model to the data

Assessing unidimensionality

Linear factor analyses: The first four eigenvalues (λ) from the interitem correlation matrices (with tetrachoric and phi correlations) and the percentages of common variance (Com.%) and of total variance (Tot.%) accounted for are displayed in Tables 13 and 14, respectively, for principal-components analysis and principal-axes factoring.

As can be seen from Tables 13 and 14, factor analyses using both tetrachoric and phi correlations found results consistent with the idea that a first dominant factor underlie the present data. Clearly, most of the variance in the data is associated with the first factor. Thus, all the factors beyond the first factor can reasonably be considered as minor factors, based on the relative sizes of the eigenvalues and the variance accounted for by the factors.

The graphical technique of the scree plot was also performed and its result was consistent with the existence of a first dominant factor.

Table 13. First Four Eigenvalues and Percentages of Variance Accounted for (Principal-Components Analysis)

	Tetrachoric Correlations			Phi Correlations		
Component	λ	Com. %	Tot. %	λ	Com. %	Tot. %
1	14.59	71.54	45.60	9.73	66.69	30.41
2	2.28	11.19	7.13	1.86	12.74	5.81
3	2.02	9.90	6.31	1.67	11.47	5.23
4	1.50	7.38	4.70	1.33	9.13	4.17

Table 14. First Four Eigenvalues and Percentages of Variance Accounted for (Principal-Axes Factoring)

	Tetrachoric Correlations			Phi Correlations		
Factor	λ	Com. %	Tot. %	λ	Com. %	Tot. %
1	14.23	75.10	74.52	9.16	74.62	74.39
2	1.93	10.19	10.11	1.30	10.56	10.53
3	1.70	8.10	8.89	1.11	9.06	9.03
4	1.09	5.74	5.70	.71	5.77	5.74

Table 15 shows the results of the parallel analysis based on 40 samples of random data.

Since the fourth eigenvalue from the random data was larger than that for the observed data, the results of the parallel analysis indicated a *maximum* of three factors. However, if one were to plot the two sets of eigenvalues, the plots obtained would be similar, except for the first eigenvalue for the real data, since this eigenvalue was substantially larger than its counterpart in the random data.

Table 15. First Four Eigenvalues (λ) for observed and random data from Parallel Analysis

Sequence Number	λ For Observed Data	Mean λ For Random Data
1	9.17	1.42
2	1.31	1.23
3	1.13	1.09
4	.72	1.00

The results of the linear factor analyses were explored further. The residual correlation matrices after one, two, three, and four factors were extracted by principal-axis analysis, were examined. The percentages of residual correlations that were greater than 0.05 (in absolute value) were respectively, 53%, 47%, 39%, and 34% after one, two, three, and four factors were extracted by PAF. This shows that much is not gained by retaining more than one factor.

Nonlinear factor analyses: These analyses focused on the examination of residuals.

NOFA results: Table 15 provides a summary of the residuals (in absolute values) after one- and two-factor NLFA models were fitted to the data. NOFA is an implementation of Etezadi-Amoli and McDonald's (1988) polynomial approximation approach to NFLA.

Table 16. Description of Residual Matrices After Fitting Linear and Non-linear Factor Analysis Models

	Goodness of Fit Index			
LFA Models	$\overline{r}_{ij}$	$s(r_{ij})$	$\lvert r_{ij}\rvert$	$s(\lvert r_{ij}\rvert)$
1-factor	.005	.034	.042	.029
2-factor	-.003	.019	.022	.024
3-factor	-.001	.013	.018	.009
4-factor	.000	.006	.015	.007
NLFA Models				
1-factor, linear term	.003	.017	.021	.012
1-factor, quadratic term	.001	.011	.016	.010
1-factor, cubic term	.000	.012	.019	.009
2-factor, linear terms	.000	.009	.010	.005

As can be seen from Table 16, both the mean and standard deviation of the absolute values of the residuals associated with a one-factor NFLA model with quadratic term were smaller than the corresponding values for residuals obtained from a two-factor

LFA. This means that a one-factor NLFA can summarize the bulk of the common variance in the data.

NOHARM results: A one-dimensional model and a two-dimensional model were fitted using NOHARM. Two residual covariance matrices were obtained, one for each model. The program also gave two goodness-of-fit indexes: the sum of squares of residuals and the root mean square of residuals. Both indicate the extent of fit between the data and the fitted model. The very small size of these indexes for even the one-dimensional model provided some support to the hypothesis that there exists only one dominant dimension underlying the data.

The sums of squares of the residuals for the one-dimensional model and the two-dimensional model were used to compute De Champlain and Gessaroli's IFI. The obtained value for the IFI was .136, which means that the two-dimensional model accounts for only a very small proportion of the sum of squares of the residual covariances from the one-dimensional model. The conclusion reached by using the IFI is the same as using the other goodness-of-fit indexes provided by NOHARM. This is understandable because the IFI is based on the sums of squares of the residual covariances.

DIMTEST analysis: The computed value of Stout's T statistic was .736, and its p-value was .231. Since this p-value is larger than .05, the null hypothesis (i.e., Ho : $d_E = 1$) was not rejected. This implies that essential unidimensionality holds for the data of the present study. This conclusion is consistent with all the results foregoing analyses pointing to the existence of a first dominant factor underlying the data.

In summary, various statistical techniques were used to assess the unidimensionality assumption for the research data. All analyses converged on the conclusion that the data set was essentially unidimensional.

Many procedures were used here because to date there is no agreement among investigators about the best method(s) for assessing the unidimensionality assumption of the most popular IRT models.

Assessing the assumption of equal discrimination indices

Both the discrimination indexes and the point-biserial correlations of the items with the total test were reviewed. As can be seen from Tables 5, 6, 7, 8, and 9, there was some variability among both discrimination indexes and point-biserial correlations. It was concluded that the assumption of discrimination indices was not met by the present data. However, it was also decided to proceed with the modeling attempt since Dinero and Haertel (1977) have reported that the Rasch model was applicable in the face of varying item discriminations.

Assessing the assumption of minimal guessing

The performance of the 24 participants with single-digit scores was examined for Compare 5, Compare 6, Change 5, and Change 6 problems, which were some of the most difficult word problems used in the present study. These children were the low-ability group in this study, as revealed by their total scores on the 32 problems. Being the low-ability group, it was very likely that they would not be able to give correct solutions to the most difficult problems, except by guessing. Consistent with this reasoning almost all of these children answered incorrectly to the most difficult problems.

This implies that these children did not guess the answers to the problems. Furthermore, the instructions stressed the importance of "thinking through" each problem, and discouraged a guessing strategy. Finally, the two experimenters reported of not having noticed (as far as they could tell) guessing tendencies in the participants.

The format used in the instrument (i.e., multiple-choice) makes it difficult to be sure that guessing was not used by subjects. Indeed, this is one of the reasons there has been much debate about the workability of the Rasch model with multiple choice items, in the IRT literature (e.g., Andrich, 1989; Divgi, 1986, 1989; Henning, 1989). In the present study, guessing was not viewed as an unavoidable consequence of the multiple choice item format. Rather, whether guessing was operating or not was considered an empirical question here. It was also thought that instructions given to participants discouraged guessing. Based on all the foregoing, it was concluded that the assumption of minimal guessing was met in the present data set.

Assessing the assumption of nonspeeded test administration

Based on the fact that children were given enough time to solve the problems and that every child completed all the problems, it was at once concluded that this assumption was fulfilled.

Item calibration using BICAL: The computer program BICAL (Wright & Stone, 1979) was used to estimate the Rasch item difficulty parameters. Some (mis)fit statistics were also obtained through this analysis. The item difficulties and their standard errors, as estimated by the unconditional maximum likelihood estimation procedure implemented in BICAL, are presented in Table 17.

Table 17. Rasch item difficulties (β) and their standard errors (SE) as estimated by BICAL

Item	β estimates	SE	Fit Mean Square
1	-1.759	.274	.93
2	.360	.222	1.22
3	-.119	.224	.89
4	.360	.222	1.11
5	-.516	.229	1.15
6	-1.291	.251	.94
7	.841	.224	.86
8	.026	.223	.81
9	.988	.226	.90
10	-.071	.224	.74
11	1.138	.228	.91
12	-.022	.223	.83
13	.792	.224	.99
14	-.567	.230	.84
15	-.777	.235	1.48
16	.696	.223	.66
17	-2.075	.295	1.25
18	-.671	.233	1.33
19	-.217	.225	.94
20	.504	.222	.68
21	.122	.222	.77
22	-.365	.227	.69
23	.841	.224	1.30
24	-.071	.224	1.20
25	.265	.222	.84
26	.551	.222	1.12
27	-.997	.241	.89
28	-.567	.230	1.23
29	.696	.223	1.15
30	.988	.226	1.28
31	-.022	.223	.95
32	.939	.225	.91

As can be seen from the above table, the precision of the difficulty estimates is uniform across items, as the standard errors have nearly equivalent values. The fit mean square (for items) is an item fit statistic. It is simply the mean squared residual averaged over persons. It will be large for an item if there are too many high ability persons who answered the item incorrectly and/or too many low ability persons who answered it correctly. The expected value of the fit mean square for an item is 1. The fit mean square values for misfitting items depart considerably from the expected value of 1. Table 17 shows that the fit of most of the items is acceptable.

To summarize, it was concluded that for the data of the present study there was no evidence that the various assumptions of the Rasch model, except the equal discrimination assumption, were violated. The next step of the modeling attempt was then undertaken.

3. Applying the LLTM to the data: Using LINLOG, the weight matrix Q and the participants' scores on the 32 word problems were used as input to the program.

Table 18 presents the parameter estimates of the six η's and their standard errors based on data from the whole sample.

Table 18. η_k **Parameter Estimates and their Standard Errors (SE)**

Complexity Factor	η_k Estimates	SE
Part-whole	.63017	.09519
Double-role counters	-.41541	.12853
Re-representation	-.51696	.16167
Comparative terms	.11461	.27305
Action cues	-.12739	.25047
Language consistency	.78152	.09716
Mean	.07775	.16768
Standard deviation	.49003	.07042

Note that these estimates are expressed on the Rasch model original metric (see Fischer, 1983). As can be seen in Table 18, three of the six η estimates were positive, ranging from .115 to .782. Hence, the complexity factors corresponding to these estimates (i.e., language consistency, part-whole, and comparative terms) were the most difficult for the participants in the present study. The complexity factor double-role counters had a negative relationship to item difficulty, as indicated by the negative value of the corresponding parameter (η = -.415). That means that the participants in the present study found that problems involving single-role counters were more difficult than those involving double-role counters. Re-representation was also negatively related to item difficulty (η = -.517). The standard errors of the η's ranged from .095 to .273, with a mean of .168 and a standard deviation of .070, indicating an overall adequate precision of the estimates.

Table 19 shows the LLTM difficulty estimates (β^*) and the Rasch model difficulty estimates (β) for the 32 items. The Rasch item difficulties displayed in Table 19 were estimated by the conditional maximum likelihood procedure implemented in LINLOG. In comparing these β parameter estimates with the ones in Table 17, which were estimated by the unconditional maximum likelihood method used in BICAL, it became quickly apparent that, for most practical purposes, the estimates could be regarded as identical.

One way of evaluating the adequacy of the hypothesized model for the whole sample and for each grade level is to assess the similarity or relationship between the two sets of item difficulty estimates. This was done by computing Pearson correlation coefficients (r). For the whole sample, the Pearson correlation coefficient between the β^*'s and the β's was statistically significant, r (30) = .74, $p < .01$, indicating that there was a fairly strong positive association between the two sets of parameter estimates. All three correlation coefficients computed using the data at each grade level were also significant. For Grade 1, r (30) = .78, $p < .01$; for Grade 2, r (30) = .67, $p < .01$; and for Grade 3, r (30) = .58, $p < .01$. These results mean that the β^*'s, which are based on the formal model, can be used to predict the Rasch item difficulties, using the data for the whole sample, and the data for each grade level.

To further evaluate the adequacy of the cognitive model based on the six complexity factors (full cognitive model), the log-likelihood χ^2 test was performed. Table 20 displays the log-likelihoods for the LLTM and the Rasch model for the total group and for the three grades.

Table 19. Normalized LLTM (β^*) and Rasch Model (β) Estimates

Item	β^*	β
1	-.57697	-1.75499
2	.31772	.36047
3	-.21086	-.11934
4	.31772	.36047
5	-.99238	-.51687
6	-.57697	-1.28934
7	1.07668	.83987
8	.44654	.02579
9	-.33498	.98642
10	.44654	-.07079
11	1.07668	1.13514
12	-.33498	-.02242
13	.20454	.79141
14	-.36224	-.56808
15	-.57697	-.77756
16	.05316	.69497
17	-.57697	-2.06310
18	-.36224	-.67181
19	-.23485	-.21707
20	.05316	.50347
21	.05316	.12181
22	-.57697	-.36554
23	.20454	.83987
24	-.57697	-.07079
25	.20454	.26515
26	.20454	.55122
27	-.57697	-.99657
28	-.57697	-.56808
29	.31772	.69497
30	1.07668	.98642
31	.31772	-.02242
32	1.07668	.93735

Table 20. Log-likelihoods (full model)

Group	LLTM	Rasch Model	χ^2	DF
1	- 597.58	- 565.62	63.90	26
2	- 610.51	- 584.10	52.82	26
3	- 520.42	- 496.81	47.21	26
Total	-1728.51	-1646.53	163.94	26

All the χ^2 values were significant at the .01 level, indicating that for the total group and for each grade level, the fits were different for the LLTM and the Rasch model. Since the likelihood ratio test gives an overall measure of goodness-of-fit of the model under consideration, these results imply that the LLTM and the Rasch model accounted for item difficulties differently, for the total group and at each grade level. In other words, the full cognitive model consisting of six complexity factors did not provide item difficulties (the β^* estimates) close enough to the Rasch item difficulties (the β estimates). This conclusion, however, should be qualified by: (i) the fact that the significance of the test depends on sample size. (No matter how good its fit to the data, any hypothesized model can be rejected, given a sufficiently large sample size.) (ii) the existence of fairly high positive correlations between the LLTM item difficulties and the Rasch item difficulties.

In conclusion, although the fit of the model to the data was quite good, to "reproduce" Rasch item difficulty parameters more closely, other complexity factors, not included in the present model, would be needed.

Recall that the full cognitive model under consideration here, consisting of the six complexity factors, contains two submodels of children's performance on arithmetic word problems, namely, the logico-mathematical submodel and the linguistic submodel. The first submodel consists of three complexity factors, part-whole, double-role counters, and re-representation. The second submodel also consists of three complexity factors, comparative terms, action cues, and language consistency. Each submodel and the full model can be viewed as (hierarchically) nested models.

The goodness-of-fit of the logico-mathematical submodel was examined for the whole sample and for each grade level. For the whole sample, the correlation between item difficulties obtained under LLTM and those based on the Rasch model was significant, $r\ (30) = .51$, $p < .01$. The correlations for all three grade levels were also significant: For Grade 1, $r\ (30) = .53$, $p < .01$; for Grade 2, $r\ (30) = .43$, $p < .05$; and for Grade 3, $r\ (30) = .43$, $p < .05$.

The linguistic submodel also was evaluated for fit by inspecting the Pearson correlations between the two sets of item difficulties. For the whole sample, $r\ (30) = .65$, $p < .01$. At each grade level, the correlation was also significant: For Grade 1, $r\ (30) = .70$, $p < .01$; for Grade 2, $r\ (30) = .61$, $p < .01$; and for Grade 3, $r\ (30) = .48$, $p < .01$.

However, there were also substantial decreases in the sizes of the correlations, moving from the full cognitive model to the each of the two submodels. More specifically, if the logico-mathematical submodel was considered rather than the full cognitive model, for the whole sample, the correlation dropped from .74 to .51; for Grade 1, from .78 to .53; for Grade 2, from .67 to .43; and for Grade 3, from .58 to .43. Moving from the full cognitive model to the linguistic submodel, the correlation for the whole sample decreased from .74 to .65; for Grade 1, from .78 to .70; for Grade 2, from .67 to .61; and for Grade 3, from .58 to .48. Decreases in the correlations were smaller when the linguistic submodel was substituted for the full cognitive model than when the logico-mathematical submodel was considered instead of the full cognitive model.

In the present research, the logico-mathematical submodel and the full cognitive model were nested models. Hence, the log-likelihood χ^2 was used to further compare the fits of the two models. For the whole sample, a significant difference was found between the two models, χ^2 (3) = 105.76, $p < .001$. The χ^2 test results were also significant for Grade 1, χ^2 (3) = 60.95, $p < .001$; and for Grade 2, χ^2 (3) = 38.31, $p < .001$; but not for Grade 3, χ^2 (3) = 6.51, $p > .05$.

Likewise, the linguistic submodel and the full cognitive model were nested models. Based on the whole sample, the difference between the two models was significant, χ^2 (3) =47.86, $p < .001$. The log-likelihood χ^2 results were also significant for Grade 1, χ^2 (3) = 26.86, $p < .001$, for Grade 2, χ^2 (3) = 9.37, $p < .05$, and for Grade 3, χ^2 (3) = 12.41, $p < .01$.

These results indicate that for the whole sample and for Grades 1 and 2, the full cognitive model and the logico-mathematical submodel accounted for item difficulties differently. In other words, for the whole sample and for Grades 1 and 2, the predictive power of the estimates was significantly improved by the addition of the linguistic submodel to the logico-mathematical submodel. This was consistent with the substantial decreases in the correlations for the whole sample and for Grades 1 and 2 when the logico-mathematical submodel was substituted for the full cognitive model.

For Grade 3, the fit of the model did not significantly deteriorate when the logico-mathematical submodel was used in lieu of the full cognitive model. This was also reflected in the relatively small decrease in the correlation (from .58 to .43). This means that the three linguistic complexity factors did not contribute significantly in accounting for the Rasch item difficulties for Grade 3. But, note that the fits of both the full model and the logico-mathematical submodel were generally poorer for Grade 3 than for the whole sample and for Grades 1 and 2.

When the linguistic submodel was used instead of the full cognitive model, the significance of the log-likelihood χ^2 tests for the whole sample and for each grade level imply that the two models accounted for Rasch item difficulties differently. In other words, for the whole sample and for each grade level, the addition of the logico-mathematical submodel to the linguistic submodel significantly improved the prediction of Rasch item difficulties.

To summarize, for the whole sample and for each grade level, (i) the logico-mathematical submodel provided an acceptable fit to the data; (ii) the linguistic submodel also provided an acceptable fit to the data; (iii) furthermore, for the whole sample and for Grades 1 and 2, but not for Grade 3, the logico-mathematical submodel contributed significantly to the predictive power of the LLTM estimates, beyond that provided by the linguistic submodel; (iv) for the whole sample and for each grade level, the linguistic submodel significantly improved the prediction of Rasch item difficulties, when it was added to the logico-mathematical submodel.

The contributions of specific complexity factors to item difficulties were also examined for the whole sample and for each grade level. Table 21 displays the parameter estimates, standard errors, and t-values for the full model using data from the whole sample. Tables 22, 23, and 24 present the parameter estimates, standard errors, and t-values, respectively, for Grades 1, 2, and 3, for the same model. The same parameter estimates based on the logico-mathematical submodel (Submodel 1) are shown in Table 25 for the whole sample, and in Tables 26, 27, and 28, respectively, for Grades 1, 2, and 3. For the linguistic submodel (Submodel 2), the parameter estimates are presented in Table 29 for the whole sample, and in Tables 30, 31, and 32, respectively, for Grades 1, 2, and 3.

For the full model, considering the whole sample as well as each grade level, the two complexity factors that were the strongest contributors to item difficulty were part-whole and language consistency. Double-role counters and re-representation had a negative relationship to item difficulty, as evidenced by their negative parameter estimates. Action cues and comparative terms did not contribute to item difficulty. This might be due to the fact that the estimates for action cues and comparative terms had very large standard errors, indicating that the parameters for these two complexity factors were not estimated with precision.

Table 21. Full modei η estimates, standard errrors, and t-values (whole sample)

Complexity factor	η_k	SE	t-value
Part-whole	.63	.10	6.30**
Double-role counters	-.42	.13	-3.20**
Re-representation	-.52	.16	-3.30**
Comparative terms	.11	.27	.41
Action cues	-.13	.25	-.52
Language consistency	.78	.10	7.80**

** $p < .01$

Table 22. Full model η estimates, standard errors, and t-values (Grade 1)

Complexity factor	η_k	SE	t-value
Part-whole	.83	.17	4.90**
Double-role counters	-.36	.20	-1.80
Re-representation	-.52	.29	-1.80
Comparative terms	.64	.49	1.30
Action cues	.00	.41	.00
Language consistency	.92	.16	5.80**

** $p < .01$

Table 23. Full model η estimates, standard errors, and t-values (Grade 2)

Complexity factor	η_k	SE	t-value
Part-whole	.52	.16	3.25**
Double-role counters	-.52	.22	-2.36*
Re-representation	-.66	.27	-2.44*
Comparative terms	.15	.46	.33
Action cues	-.02	.43	-.05
Language consistency	.82	.16	5.13**

* $p < .05$, ** $p < .01$

Table 24. Full model η estimates, standard errors, and t-values (Grade 3)

Complexity factor	η_k	SE	t-value
Part-whole	.59	.17	3.47**
Double-role counters	-.39	.25	-1.56
Re-representation	-.45	.29	-1.55
Comparative terms	-.39	.50	-.78
Action cues	-.37	.46	-.80
Language consistency	.58	.18	3.22**

** $p < .01$

Table 25. Submodel 1 η estimates, standard errors, and t-values (Whole sample)

Complexity factor	η_k	SE	t-value
Part-whole	.73	.08	12.5**
Double-role counters	-.68	.11	-6.2**
Re-representation	-.23	.13	-1.77

** $p < .01$

Table 26. Submodel 1 η estimates, standard errors, and t-values (Grade 1)

Complexity factor	η_k	SE	t-value
Part-whole	.93	.15	6.20**
Double-role counters	-.76	.18	-4.22**
Re-representation	-.24	.23	-1.04

** $p < .01$

Table 27. Submodel 1 η estimates, standard errors, and t-values (Grade 2)

Complexity factor	η_k	SE	t-value
Part-whole	.60	.15	4.00**
Double-role counters	-.79	.19	-4.16**
Re-representation	-.31	.22	-1.41

** $p < .01$

Table 28. Submodel 1 η estimates, standard errors, and t-values (Grade 3)

Complexity factor	η_k	SE	t-value
Part-whole	.64	.16	4.00**
Double-role counters	-.48	.21	-2.29*
Re-representation	-.16	.23	-.69

* $p < .05$; ** $p < .01$

Table 29. Submodel 2 η estimates, standard errors, and t-values (Whole sample)

Complexity factor	η_k	SE	t-value
Comparative terms	.22	.24	.92
Action cues	-.24	.22	-1.09
Language consistency	.76	.08	9.5**

** $p < .01$

Table 30. Submodel 2 η estimates, standard errors, and t-values (Grade 1)

Complexity factor	η_k	SE	t-value
Comparative terms	.51	.39	1.31
Action cues	-.30	.36	.83
Language consistency	.92	.14	6.57**

** $p < .01$

Table 31. Submodel 2 η estimates, standard errors, and t-values (Grade 2)

Complexity factor	η_k	SE	t-value
Comparative terms	.46	.41	1.12
Action cues	.01	.38	.03
Language consistency	.73	.14	5.21**

** $p < .01$

Table 32. Submodel 2 η estimates, standard errors, and t-values (Grade 3)

Complexity factor	η_k	SE	t-value
Comparative terms	-.27	.43	-.63
Action cues	.47	.41	-1.15
Language consistency	.57	.15	3.8**

** $p < .01$

For the logico-mathematical submodel, for the whole sample and for each grade level, part-whole contributed most to item difficulty. Double-role counters had again a negative relationship to item difficulty. For the linguistic submodel, for the whole sample and for each grade level, only language consistency contributed to item difficulty. The other two complexity factors (i.e., comparative terms and action cues) had rather large standard errors. This lack of precision in their estimates was the major reason for the lack of effectiveness of these two complexity factors.

To summarize, in considering item difficulty as conceived by the cognitive model used in the LLTM, three of the complexity factors made significant contributions to item difficulty, namely, the part-whole schema, double-role counters, and language consistency. Note that the first two complexity factors were related to conceptual knowledge, whereas the third factor was related to linguistic knowledge. The part-whole schema was found to be a knowledge aspect making some problems more difficult than others, a finding consistent with the RGH model. The contributions of the two other complexity factors were in a direction opposite to what was expected. More specifically, for the logico-mathematical submodel, the complexity factor double-role counters had an opposite contribution at all grade levels. For the full cognitive model, double-role counters and re-representation had an opposite contribution, but only at Grade 2.

To provide converging evidence, the observed item difficulties were changed to logits using the logit transformation, ln (p / 1 - p). Then, these logits were regressed on the Q matrix (the matrix of item scores on the complexity factors). The full model and the two submodels were evaluated separately at each grade level. The change in R^2

from one submodel to the full model was tested for significance as a way of evaluating the contribution of a submodel entered at the second step of the regression procedure. This was done using incremental F ratios (F_{inc}). Tables 33, 34, and 35 present the results of this analysis, respectively, for Grades 1, 2, and 3.

Table 33. Regression analysis results for Grade 1

Model	R^2	F	Sig. F	F_{inc}	Sig. F_{inc}
Full cognitive model (f)	.60	6.28	.0004		
Logico-math. submodel (1)	.23	2.83	.0562	.92	> .10
Linguistic submodel (2)	.53	10.36	.0001	.54	> .10

Table 34. Regression analysis results for Grade 2

Model	R^2	F	Sig. F	F_{inc}	Sig. F_{inc}
Full cognitive model (f)	.54	4.94	.0019		
Logico-math. submodel (1)	.24	2.89	.0532	.50	> .10
Linguistic submodel (2)	.42	6.87	.0013	.20	> .10

Table 35. Regression analysis results for Grade 3

Model	R^2	F	Sig. F	F_{inc}	Sig. F_{inc}
Full cognitive model (f)	.41	2.92	.0269		
Logico-math. submodel (1)	.27	3.46	.0295	.20	> .10
Linguistic submodel (2)	.21	2.54	.0765	.29	> .10

As can be seen, the full cognitive model led to a significant R^2 at each grade level. However, the R^2 for the logico-mathematical submodel was significant at Grade 3, but barely reached significance at Grades 1 and 2; the R^2 for the linguistic submodel was significant at Grades 1 and 2, but was not significant at Grade 3. What this means is that the following trend is clearly apparent in the data: As one moves from Grades 1 to 3, the logico-mathematical complexity factors increase in importance, whereas the linguistic complexity factors decrease. The increments in R^2 when the linguistic complexity factors were added to the logico-mathematical complexity factors were .37, .20, and .14, respectively, for Grades 1, 2, and 3. However, the increments in R^2 when the logico-mathematical complexity factors were added to the linguistic complexity factors were .07, .12, and .20, respectively, for Grades 1, 2, and 3. Note that these trends were only suggestive since they failed to be reliable, as shown by the non-significance of all the incremental F ratios displayed in Tables 33, 34, and 35.

The regression analysis also showed the following significant complexity factors in terms of their contribution to the improvement of the fit. For the linguistic submodel, language consistency was a significant complexity factor at all grade levels. For the logico-mathematical submodel, part-whole was the only significant complexity factor at all grade levels. For the full cognitive model, part-whole and language consistency were the significant contributing complexity factors at all grade levels.

The contributions of two other complexity factors were in a direction opposite to what was expected. More specifically, for the full cognitive model, double-role counters and re-representation had an opposite contribution, but only at Grade 2. For the logico-mathematical submodel, the complexity factor double-role counters had an opposite contribution at all grade levels. Clearly, these results are consistent with the results of the LLTM analysis presented earlier.

The results provided some support to the RGH model in that they were consistent with the notion that the requirement that children possess knowledge of part-whole relations makes some arithmetic word problems more difficult than others. The negative relationship between double-role counters and children's performance on the problems was in contradiction with the BL model according to which problems involving double-role counters are more difficult than problems involving single-role counters. The notion that language that is inconsistent with the arithmetic operation required to solve a given problem (e. g., "altogether" in a subtraction problem, "fewer" or "less" in an addition problem) makes the problem more difficult for children received some empirical support, because the complexity factor of language consistency contributed significantly to the difficulty of the problems. Hence, some aspects of both conceptual knowledge and linguistic knowledge were found to be factors influencing the difficulty of addition and subtraction word problems for children who participated in the present study.

V. DISCUSSION

The present study was motivated by the need for more research dealing with the role of conceptual knowledge and linguistic knowledge in children's performance on addition and subtraction word problems. In an attempt to provide answers to the question of the relative difficulty of these problems, many researchers have adopted a competence model of problem solving. The emphasis on conceptual knowledge or the understanding of the semantic relations in problem texts led to the notion that a competent problem solver possesses schemata that represent these semantic relations and connect them to solution sequences. Consequently, problem-solving skill is believed to improve in direct relation with children's capabilities of representation.

The acquisition of semantic schemata depends on the acquisition of knowledge concerning logical set relations in general, and part-whole relations in particular. Acquisition of knowledge concerning part-whole relations allows a problem solver to represent word problems as relations among parts and wholes. Other aspects of conceptual knowledge, such as the understanding that a given item can be part of two sets simultaneously, and the understanding that problems can be re-represented. In the literature on children's performance on arithmetic word problems, the emphasis on conceptual knowledge has been referred to as the logico-mathematical view.

In the present research, a second position concerning the locus of improvements in children's skills in solving arithmetic word problems, namely, the linguistic development view was contrasted with the logico-mathematical view. The linguistic development view holds that linguistic sophistication is as important as logico-mathematical knowledge in children's performance on arithmetic word problems. Three aspects of linguistic knowledge were considered here, namely, children's understanding of comparative terms, of action cues, and of inconsistent language in problem texts.

A psychometric modeling approach was taken to further investigate the role of logico-mathematical knowledge and of linguistic knowledge in children's performance on arithmetic word problems.

In this part of the thesis, the results are summarized, the significance and implications of the study are discussed, and the directions for future studies are outlined.

The modeling approach used here sought to explain the difficulty of addition and subtraction word problems by referring to three aspects of logico-mathematical knowledge and three aspects of linguistic knowledge.

First, the data were examined using several procedures designed for the assessment of model-data fit. The assumptions of the model being used (i.e., the LLTM), which is a constrained Rasch model, were evaluated using the research data. The assessment of traditional and essential unidimensionality was quite extensive, as it involved various procedures used in other studies dealing with the assessment of unidimensionality, or recommended by measurement experts.

As in previous studies (e.g., Hambleton & Rovinelli, 1986), procedures based on fitting a one-factor nonlinear model and examining the residual correlations were most revealing, as were Stout's (1987) T statistic of essential unidimensionality and De Champlain and Gessaroli's incremental fit index. The use of these procedures is recommended for future investigations. Based on these procedures, the conclusion that a dominant first factor underlie the present data was rather compelling.

However, this was not quite the case for the assessment of the other assumptions of the model under consideration, namely, the assumptions of equal discrimination indices and of minimal guessing, although a case was also made, on both empirical and logical grounds, about their acceptability.

The end result of most of the preliminary analyses and argumentation was the conclusion that it was not unreasonable to consider that the data provided enough evidence for fit to the Rasch model (or more precisely, that there was no obvious evidence of model-data misfit).

The evaluation of the psychometric model (i.e., the LLTM) applied to the data was a major part of the data analysis in the present study. One important finding of that analysis was that the linear combination of complexity factors based on the three aspects of conceptual knowledge (i.e., part-whole, double-role counters, and re-representation) and the three aspects of linguistic knowledge (i.e., comparative terms, action cues, and consistency of language) was adequate to account for the item difficulties based on the Rasch model. This result was true for the whole sample and for each grade level, as relatively high positive correlations were found between the LLTM item difficulties and the Rasch item difficulties.

However, it was also found that the LLTM and the Rasch model did not account for item difficulty exactly the same way. Since all the log-likelihood χ^2 tests were significant for the whole sample and for each grade level, the item difficulties based on the LLTM were different from those based on the Rasch model, in each case. This result was an indication that the full cognitive model consisting of all six complexity factors was not sufficient as a basis for a model that could provide item difficulties (the β^* estimates) close enough to the Rasch item difficulties. In other words, to make possible to reproduce Rasch item difficulties more closely, that is, to improve the predictive power of the η parameter estimates, other aspects of conceptual knowledge, linguistic knowledge, and/or any other aspects, not included in the present model, would be necessary.

Both the logico-mathematical and the linguistic submodels were also adequate models accounting for the difficulty of arithmetic word problems, but necessarily less so than the full cognitive model. For the whole sample and for Grades 1 and 2, the fit of the submodel under consideration was significantly improved by adding the other submodel. Hence, for children in Grades 1 and 2, complexity factors related to logico-mathematical and linguistic knowledge made some problems more difficult than others.

For Grade 3, no improvement in the predictive power was gained by moving from the logico-mathematical submodel to the full cognitive model, that is, by adding the linguistic submodel to the logico-mathematical submodel. Thus, for Grade 3, the three linguistic complexity factors used in the present study did not add significantly to the predictive power of the three logico-mathematical complexity variables.

Among the three logico-mathematical complexity variables two were found to influence item difficulty, part-whole schema and double-role counters. As already mentioned, the part-whole schema has been viewed as the hallmark of the most advanced level of arithmetic word problem-solving competence, according to RGH model (e.g., Riley & Greeno, 1988). The present results are consistent with the view that knowledge of part-whole relationships is a knowledge requirement making some arithmetic word problems more difficult than others, for children in the early grades. Contrary to what was expected on the basis of the BL model, single-role rather than double-role counters, made some problems more difficult than others for the partcipants in the present study.

Among the three linguistic complexity variables, only consistency of language was a reliable factor influencing the difficulty of arithmetic word problems. Problems in which there were words that children found inconsistent with the operation required to solve them were more difficult than others. Comparative terms as a linguistic complexity factor did not contribute to the difficulty of the problems, as revealed by the LLTM analysis. However, it was noted above that compare problems were more difficult than the other types of problems. This is only an apparent contradiction. Note that not all compare problems are more difficult than the other types of problems. Compare problems known to be very difficult for children are usually those in which the comparative terms are inconsistent with the operation called for in the solution of the problems (i.e., "more" in a subtraction problem, "less" in an addition problem). Half of the compare problems used in the present study involved inconsistent language. In other words, compare problems are not difficult only because they contain comparative terms; but rather when these comparative terms occur with words that are seemingly inconsistent with them, in children's view. Recall that consistency of language was a complexity factor found to contribute significantly to the difficulty of the problems in this study. In fact, a consistency hypothesis has been advanced by Lewis and Mayer (1987) in an attempt to account for children's relatively poor performance on those problems. The present results are consistent with that hypothesis.

In conclusion, both logico-mathematical and linguistic knowledge were found to be complexity factors influencing the difficulty of addition and subtraction word problems for children in Grades 1, 2, and 3. But not all the aspects of these two types of knowledge, as hypothesized by two current theories of children's performance on arithmetic word problems, contributed significantly in influencing problem difficulty. The fact that part-whole and consistency of language had significant contributions in determining the difficulty of the problems has clear educational implications. First, for some problems, helping children understand part-whole relationships among sets may have a facilitating effect in children's development of skill in solving arithmetic word problems. Second, words in problem texts or statements that children view as inconsistent with the arithmetic operation required to solve the problems need to be avoided, especially for the younger children, as these are the ones more prone to rely on the rule that, for example, "altogether" and "more" always call for addition, "less" for subtraction.

As illustrated in the present research, the LLTM is a powerful modeling tool that can be useful in investigating the structure of cognitive tasks. Here task features were used as complexity factors. However, complexity factors are not limited to task features in applications of the LLTM. For instance, the processing operations used by problem solvers are always good candidates for complexity factors. Indeed, the use of processing operations together with task features may improve future LLTM analyses, which are applicable to the investigation of many research questions, especially those dealing with performance on cognitive tasks. Future LLTM analyses will also benefit from the use of larger sample sizes. An obvious limitation of the present study is the small sample size used. Hence, more confidence will be placed in the present results when they are replicated using far larger sample sizes. Another suggestion for further research is that continued use of the LINLOG computer program in LLTM analyses will benefit from studies examining the standard errors supplied by the program; bootstrap or jacknife procedures are suggested for these studies.

REFERENCES

Ackerman, T. (1989). Unidimensional IRT calibrations of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113-127.

Andrich, D. (1989). Statistical reasoning in psychometric models and educational measurement. *Journal of Educational Measurement, 26,* 81-90.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37-48.

Arnett, L. D. (1905). Counting and adding. *American Journal of Psychology, 16,* 327-336.

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review, 2,* 213-236.

Ashcraft, M. H. & Fierman, B. A. (1982). Mental addition in third, fourth, and sixth graders. *Journal of Experimental Child Psychology, 33,* 216-234.

Ashcraft, M. M., Fierman, B. A., & Bartolotta, R. (1984). The production and verification tasks in mental addition: An empirical comparison. *Developmental Review, 3,* 157-170.

Assessment Systems Corporation (1988). *User's manual for the Micro-CAT Testing System, Version 3.* St. Paul, MN: Author.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameters logistic models. *Applied Psychological Measurement, 11,* 111-141.

Baroody, A. J. (1984). The development of procedural knowledge: An alternative explanation for chronometric trends in mental arithmetic. *Developmental Review, 3,* 225-230.

Batley, R. M., & Boss, M. W. (1993). The effects on parameter estimation of correlated dimensions and a distribution-restricted trait in a multidimensional item response model. *Applied Psychological Measurement*, 17, 131-141.

Behr, M. J., Lesh, R., Post, T. R., & Silver, E. A. (1983). Rational number concepts. In R. Lesh & M. Landau (Eds.), *Acquisition of mathematics concepts and processes* (pp. 91-126). New York: Academic Press.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), *Applications of item response theory.* Vancouver, BC: Educational Research Institute of British Columbia.

Birnbaum, A. (1968). Test scores, sufficient statistics, and the information structures of tests. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R. D., & Mislevy, R. J. (1981). An item response curve model for matrix sampling data: The California Grade Three assessment. In D. Carlson (Ed.), *New directions for testing and measurement: Testing in the states.* San Francisco: Jossey-Bass.

Brainerd, C. J. (1983). Young children's mental arithmetic errors: A working-memory analysis. *Child Development, 54,* 812-830.

Briars, D. J., & Larkin, J. H. (1984). An integrated model of skill in solving elementary word problems. *Cognition and Instruction, 1,* 245-296.

Briars, D., & Siegler, R. S. (1986). A featural analysis of preschoolers' counting knowledge. *Developmental Psychology, 20,* 607-618.

Browne, C. E. (1906). The psychology of simple arithmetical processes: A study of certain habits of attention and association. *American Journal of Psychology, 17,* 1-37.

Brownell, W. A. (1928). *The development of children's number ideas in the primary grades.* (Supplementary Educational Monograph no. 35). Chicago: University of Chicago Press.

Brownell, W. A. (1947). An experiment in "borrowing" in third-grade arithmetic. *Journal of Educational Research, 41,* 161-171.

Brownell, W. A., & Stretch, L. B. (1931). *The effects of unfamiliar settings on problem setting.* Durham, NC: Duke University.

Bruner, J. S. (1960). *The process of education.* New York: Vintage Books.

Buswell, G. T., & Judd, H. (1925). *Summary of educational investigations relating to arithmetic* (Supplementary Educational Monograph no. 27). Chicago: University of Chicago Press.

Campbell, J. I. D., & Graham, D. J. (1985). Mental multiplication skills: Structure, process, and acquisition. *Canadian Journal of Psychology, 39,* 338-366.

Carey, S. (1985). Are children fundamentally different kinds of thinkers and learners than adults? In S. F. Chipman, J. W. Segal, & R. Glaser Eds.), *Thinking and learning skills: Current research and open questions* (Vol. 2). Hillsdale, NJ: Erlbaum.

Carlson, J. E. (1987). Multidimensional item response theory estimation: A computer program (ACT Research Report no. 87-19). Iowa City, IA: American College Testing Program.

Carpenter, T. P. (1985). Learning to add and subtract. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives.* Hillsdale, NJ: Erlbaum.

Carpenter, T. P., & Moser, J. M. (1982). The development of addition and subtraction problem solving skills. In T. P. Carpenter, J. M. Moser, & T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 9-24). Hillsdale, NJ: Erlbaum.

Carpenter, T. P., & Moser, J. M. (1983). The acquisition of addition and subtraction concepts. In R. Lesh & M. Landau (Eds.), *Acquisition of mathematics concepts and processes* (pp. 7-44). New York: Academic Press.

Carpenter, T. P., Kepner, H., Corbitt, M. K., Linquist, M. M., & Reys, R. E. (1980). Results and implications of the second NAEP mathematics assessments: Elementary school. *Arithmetic Teacher, 28,*10-12.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215-281). New York: Academic Press.

Chi, M. T. H. (1978). Knowledge structures and memory development. In R. Siegler (Ed.), *Children's thinking: What develops?* (pp. 73-96). Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 7-75). Hillsdale, NJ: Erlbaum.

Chi, M. T. H., & Koeske, R. D. (1983). Network representation of a child's dinosaur knowledge. *Developmental Psychology, 19,* 29-39.

Clapp, F. L. (1974). *The number combinations: their relative difficulty and frequency of their occurrence in textbooks.* Bureau of Educational Research Bulletin no. 1, Madison, WI.

Cobb, P. (1987). An analysis of three models of early number development. *Journal for Research in Mathematics Education, 18*, 163-179.

Cota, A. A., Longman, R. S., Holden, R. R., & Fekken, G. C. (1993a). Comparing different methods for implementing parallel analysis: A practical index of accuracy. *Educational and Psychological Measurement, 53*, 865-876.

Cota, A. A., Longman, R. S., Holden, R. R., & Fekken, G. C. (1993b). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement, 53*, 585-595.

Cummins, D. D. (1991). Children's interpretation of arithmetic word problems. *Cognition and Instruction, 8*, 261-289.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*, 405-438.

Davis-Dorsey, J., Ross, S. M., & Morrison, G. R. (1991). The role of rewording and context personalization in the solving of mathematical word problems. *Journal of Educational Psychology*, 83, *61-68.*

Dean, A. L., & Malik, M. M. (1986). Representing and solving arithmetic word problems: A study of developmental interaction. *Cognition and Instruction, 3*, 211-227.

De Champlain, A., & Gessaroli, M. E. (1991, April). *Assessing test dimensionality using an index based on nonlinear factor analysis.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

De Corte, E., & Verschaffel, L. (1985). Beginning first graders' initial representation of arithmetic word problems. *The Journal of Children's Mathematical Behavior, 4,* 3-21.

De Corte, E., & Verschaffel, L. (1987). The effect of semantic structure on first graders' strategies for solving addition and subtraction word problems. *Journal for Research in Mathematics Education, 18,* 363-381.

De Corte, E., & Verschaffel, L, & Dewin, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology, 77,* 460-470.

De Corte, E., Verschaffel, L., & Pauwels, A. (1990). Influence of the semantic structure of word problems on second graders' eye movements. *Journal of Educational Psychology, 82,* 359-365.

Dimitrovsky, L., & Almy, M. (1975). Early conservation as a predictor of arithmetic achievement. *The Journal of Psychology, 91,* 65-70.

Dinero, T. E., & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement, 1,* 581-592.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23,* 283-298.

Divgi, D. R. (1989). Reply to Andrich and Henning. *Journal of Educational Measurement, 26,* 295-299.

Dodwell, P. C. (1961). Children's understanding of number concepts: Characteristics of an individual and a group test. *Canadian Journal of Psychology, 15,* 29-36.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186.

Etezadi-Amoli, J., & McDonald, R. P. (1983). A second generation nonlinear factor analysis. *Psychometrika*, *48*, 315-342.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics*. New York: Academic Press.

Embretson, S. E., Schneider, L. M., & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement, 23,* 13-32.

Fayol, M., Abdi, H., & Gombert, J. E. (1987). Arithmetic problems formulation and working memory load. *Cognition and Instruction, 4,* 187-202.

Fischer, F. E. (1990). A part-part-whole curriculum for teaching number in the kindergarten. *Journal for Research in Mathematics Education, 21,* 207-215.

Fischer, G. H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.

Fischer, G. H. (1978). Probabilistic test models and their applications. *German Journal of Psychology, 2,* 298-319.

Fischer, G. H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. *Psychometrika, 46,* 59-77.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Pschometrika*, 48, 3-26.

Fischer, G. H., & Forman, A. K. (1972). *An algorithm and a FORTRAN program for estimating the linear logistic test model* (Research Bulletin No. 24). Vienna: University of Vienna, Institute of Psychology.

Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 6,* 397-416.

Fraser, C. (1981). NOHARM: *A Fortran program for non-linear analysis by a robust method for estimating the parameters of 1-, 2-, and 3-parameter latent trait models.* Armidale, Australia: University of New England, Center for Behavioral Studies in Education.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267-269.

Freyberg, P. S. (1966). Cognitive development in Piagetian terms in relation to school attainment. *Journal of Educational Psychology, 57,* 164-168.

Fuson, K. C., & Willis, G. B. (1989). Second graders' use of schematic drawings in solving addition and subtraction word problems. *Journal of Educational Psychology, 81,* 514-520.

Gagne, R. M. (1977). *The conditions of learning* (3rd ed.). New York: Holt.

Gelman, R., & Meck, E. (1983). Preschoolers' counting: Principles before skill. *Cognition, 13,* 343-359.

Gelman, R., Meck, E., & Merkin, S. (1986). Young children's numerical competence. *Cognitive Development, 1,* 1-29.

Gibb, E. G. (1956). Children's thinking in the process of subtraction. *Journal of Experimental Education, 25,* 71-80.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist,* 39, 93-104.

Greeno, J. G. (1978). A study of problem solving. In R. Glaser (Ed.), *Advances in instructional psychology* (vol. 1). Hillsdale, NJ: Erlbaum.

Greeno, J. G. (1980). Trends in the theory of knowledge for problem solving. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research*. Hillsdale, NJ: Erlbaum.

Groen, G. J., & Parkman, J. M. (1972). A chronometric analysis of simple addition. *Psychological Review, 79,* 329-343.

Grouws, D. A. (1972). Open sentences: Some instructional considerations from research. *Arithmetic Teacher, 19,* 595-599.

Haman, M. S., & Ashcraft, M. H. (1985). Simple and complex mental addition across development. *Journal of Experimental Child Psychology*, *40*, 49-72.

Haman, M. S., & Ashcraft, M. H. (1986). Textbook presentations of the basic addition facts. *Cognition and Instruction, 3,* 173-192.

Hambleton, R. K. (1989). Principles and selected applications of items response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory.* Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of items. *Applied psychological measurement*, *10*, 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newburg Park, CA: Sage.

Haneghan, J. P. (1990). Third and fifth graders' use of multiple standards of evaluation to detect errors in word problems. *Journal of Educational Psychology, 82,* 352-358.

Harrison, D. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11,* 91-115.

Hattie, J. A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49-78.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Hays, R. D. (1987). PARALLEL: A program for performing parallel analysis. *Applied Psychological Measurement, 11, 58.*

Hebbeler, K. (1977). Young children's addition. *The Journal of children's mathematical behavior, 1,* 108-121.

Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal Educational Psychology, 84,* 76-84.

Henning, G. (1989). Does the Rasch model really work for multiple-choice items? Take another look: A response to Divgi. *Journal of Educational Measurement, 26,* 91-97.

Hiebert, J. (1982). The position of the unknown set and children's solutions of verbal arithmetic problems. *Journal for Research in Mathematics Education, 13,* 341-349.

Hiebert, J. A. (1985). *Conceptual and procedural knowledge: The case of mathematics.* Hillsdale, NJ: Erlbaum.

Hiebert, J., & Carpenter, T. P. (1980). Piagetian tasks as readiness measures in mathematical instruction: A critical review. *Educational Studies in Mathematics, 13,* 329-345.

Hiebert, J. A., & Lefevre, P. (1985). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. A. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1-27). Hillsdale, NJ: Erlbaum.

Hitch, G. J. (1978). The role of short-term working memory in mental arithmetic. *Cognitive Psychology, 10,* 302-323.

Horn, J. L. (1965). A rationale and technique for estimating the number of factors in factor analysis. *Psychometrika*, *24*, 265-267.

Hsu, T., & Yu, L. (1989). Using computers to analyze item response data. *Educational Measurement: Issues and Practice,* 21-28.

Hubert, L., & Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology, 29*, 190-241.

Hudson, T. (1983). Correspondence and numerical differences between disjoint sets. *Child Development, 54,* 84-90.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17,* 475-483.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bids in the prediction context. *Journal of Applied Psychology, 71,* 327-333.

Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, 10, 193-206.

Hydle, L. L., & Clapp, F. L. (1927). *Elements of difficulty in the interpretation of concrete problems in arithmetic.* Bureau of Educational Research Bulletin no. 9. Madison, Wis.: University of Wisconsin.

Ibarra, C. G., & Lindvall, C. M. (1982). Factors associated with the ability of kindergarten children to solve simple arithmetic story problems. *Journal of Educational Research, 75,* 149-155.

Jeffries, R., Turner, A. A., Polson, P. G., & Atwood, M. E. (1981). The processes involved in designing software. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition.* Hillsdale, NJ: Erlbaum.

Jerman, M. (1973). Problem length as a structural variable in verbal arithmeticproblems. *Educational Studies in Mathematics, 5,* 109-123.

Jerman, M., & Mirman, S. (1974). Linguistic and computational variables in problem solving in elementary mathematics. *Educational Studies in Mathematics, 6,* 3-28.

Jerman, M., & Rees, R. (1972). Predicting the relative difficulty of verbal arithmetic problems. *Educational Studies in Mathematics, 4,* 306-323.

Kail, R., & Bisanz, J. (1982). Information processing and cognitive development. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (vol. 17). New York: Academic Press.

Kameenui, E. J. & Griffin, C. C. (1989). The national crisis in verbal problem solving in mathematics: A proposal for examining the role of basal mathematics programs. *The Elementary School Journal, 89,* 575-593.

Kaufman, A. S., & Kaufman, N. L. (1972). Tests built from Piaget's and Gessell's tasks as predictors of first grade achievement. *Child Development, 43,* 521-535.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving arithmetic word problems. *Psychological Review, 92,* 109-129.

Kameenui, E. J. & Griffin, C. C. (1989). The national crisis in verbal problem solving in mathematics: A proposal for examining the role of basal mathematics programs. *The Elementary School Journal, 89,* 575-593.

Klahr, D. (1989). Information-processing approaches. In R. Vasta (Ed.), *Annals of child development* (vol. 6). Greenwich, CT: JAI Press.

Knight, F. B., & Behrens, M. S. (1928). *The learning of the 100 addition combinations and the 100 subtraction combinations.* New York: Longmans, Green and Co.

Kramer, G. A. (1933). *The effects of certain factors in the verbal arithmetic problems upon children's success in the solution.* The John Hopkins University Studies in Education no. 20. Baltimore, MD: The John Hopkins University Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Lachman, R., Lachman, J. L., & Butterfield, E. C. (1979). *Cognitive psychology and information processing: An introduction.* Hillsdale, NJ: Erlbaum.

Larkin, J. H., McDermott, H., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208,* 1335-1342.

Lautenschlager, G. J. (1989). PARANAL.TOK: A program for developing parallel analysis criteria. *Applied Psychological Measurement, 13,* 176.

Lewis, A. B. (1989). Training students to represent arithmetic word problems. *Journal of Educational Psychology, 81,* 521-531.

Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology, 79,* 363-371.

Lindvall, C. M., & Ibarra, C. G. (1980). Incorrect procedures used by primary grade pupils in solving open addition and subtraction sentences. *Journal for Research in Mathematics Education, 11*, 50-62.

Loftus, E. F., & Suppes, P. (1972). Structural variables that determine problem solving difficulty in computer-assisted instruction. *Journal of Educational Psychology, 63*, 631-642.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989a). PAM: A double-precision FORTRAN routine for the parallel analysis method in principal components analysis. *Behavior Research Methods, Instruments, & Computers, 21*,477-480.

Longman, R. S., Cota, A. A., Holden, R. R., & Fekken, G. C. (1989b). A regressin equation for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. *Multivariate Behavioral Research, 24*,59-69.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph, No. 7). Psychometric Society.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mayer, R. E. (1985). Implications of cognitive psychology for instruction in mathematical problem solving. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 123-145). Hillsdale, NJ: Erlbaum.

Mayer, R. E., Larkin, J. H., & Kadane, J. B. (1984). A cognitive analysis of mathematical problem solving. In R. J. Sternberg (ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 231-273). Hillsdale, NJ: Erlbaum.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R. P. (1983). Linear versus nonlinear models in item-response theory. *Applied psychological Measurement*, *6*, 379-396.

McKinley, R. L. (1989). *Confirmatory analysis of text structure using multidimensional item response theory*. Princeton, NJ: Educational Testing Service (Research Report).

McLellan, J. A., & Dewey, J. (1895). *The psychology of number and its application to methods of teaching arithmetic*. New York: Appleton.

Medina-Diaz, M. (1993). Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement*, *17*, 117-130.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

Mislevy, R. J., & Bock, R. D. (1984). BILOG: *Maximum likelihood item analysis and test scoring with logistic models*. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Morales, R. V., Shute, V. J., & Pellegrino, J. W. (1985). Developmental differences in understanding and solving simple mathematics word problems. *Cognition and Instruction, 2*, 1-39.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Mulholland, T; Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12*, 252-284.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28,* 99-117.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 17,* 29-38.

Nandakumar, R. (1994). Assessing dimensionality of item responses - Comparison of different approaches. *Journal of Educational Measurement, 31,* 17-35.

Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait dimensionality. *Journal of Educational Statistics, 18,* 41-68.

National Council of Teachers of Mathematics (1980). *An agenda for action: Recommendations for school mathematics of the 1980's.* Reston, VA: Author.

Nesher, P., Greeno, J. G., & Riley, M. S. (1982). The development of semantic categories for addition and subtraction. *Educational Studies in Mathematics, 13,* 373-394.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Pellegrino, J. W., & Glaser, R. (1982). Analyzing aptitudes for learning: Inductive reasoning. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 2, pp. 269-345). Hillsdale, NJ: Erlbaum.

Piaget, J. (1952). *The Child's conception of number.* London: Routledge & Kegan Paul.

Polya, G. (1965). *Mathematical discovery: On understanding, learning, and teaching problem solving.* (Vol. 2). New York: Wiley.

Putnam, R. T., Lampert, M., & Peterson, P. L. (1990). Alternative perspectives on knowing mathematics in elementary schools. In C. B. Cazden (Ed.), *Review of research in education* (Vol. 16). Washington, DC: American Educational Research Association.

Rasch, G. (1960). *Probabilistic models for home intelligence and attainment.* Copenhagen: Danish Institute for Educational Research.

Rathmell, E. C. (1986). Helping children learn to solve story problems. In A. Zollman, W. Speer, & J. Meyer (Eds.), *The Fifth Mathematics Methods Conference Papers* (pp. 101-109). Bowling Green, OH: Bowling Green State University.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Reckase, M. D., & McKinley, R. L. (1985). Some latent trait theory in a multidimensional latent space. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Resnick, L. B. (1983). A developmental theory of number understanding. In H. P. Ginsberg (Ed.). *The development of mathematical thinking.* New York: Academic Press.

Resnick, L. B., & Ford, W. W. (1981). *The psychology of mathematics for instruction.* Hillsdale, NJ: Erlbaum.

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education*, 20, 8-27.

Riley, L. B., & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction, 5,* 49-101.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem solving ability in arithmetic. In H. P. Ginsberg (Ed.), *The development of mathematical thinking* (pp. 153-196). New York: Academic Press.

Romberg, T. A. (1982). An emerging paradigm for research on addition and subtraction skills. In T. P. Carpenter, J. M. Moser, & T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective*. Hillsdale, NJ: Erlbaum.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 17.*

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203-219.

Siegler, R. S. (1983). Information processing approaches to development. In P. H. Mussen (Ed.), *Handbook of child psychology* (vol. 1). New York: Wiley.

Siegler, R. S., & Klahr, D. (1982). When do children learn? The relationship between existing knowledge and the acquisition of new knowledge. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 2, pp. 121-211). Hillsdale, NJ: Erlbaum.

Siegler, R. S., & Richards, D. D. (1982). The development of intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 897-971). Cambridge, England: Cambridge University Press.

Silverstein, A. B. (1987). Note on the parallel analysis criterion for determining the number of common factors or principal components. *Psychological Reports, 61,* 351-354.

Silverstein, A. B. (1990). Update on the parallel analysis criterion for determining the number of principal components. *Psychological Reports, 67,* 511-514.

Simon, H. A. (1980). Problem solving and education. In D. T. Tuma, & T. Reif (Eds.), *Problem solving and education: Issues in teaching and research.* Hillsdale, NJ: Erlbaum.

Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler (Ed.), *Children's thinking: What develops?* Hillsdale, NJ: Erlbaum.

Steffe, L. P. (1970). Differential performance of first grade children when solving arithmetic addition problems. *Journal for Research in Mathematics Education, 1,* 144-161.

Steffe, L. P., & Johnson, D. C. (1971). Problem solving performances of first grade children. *Journal for Research in Mathematics Education, 2,* 50-64.

Stern, E. (1993). What makes certain arithmetic word problems involving the comparison of sets so difficult for children? *Journal of Educational Psychology, 85,* 7-23.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1981). Testing and cognitive psychology. *American Psychologist, 36,* 1181-1189.

Stigler, J. W., Fuson, K. C., Ham, M., & Kim, M. S. (1986). An analysis of addition and subtraction word problems in American and Soviet elementary mathematics textbooks. *Cognition and Instruction, 3,* 153-171.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293-325.

Suppes, P., Loftus, E. F., & Jerman, M. (1969). Problem solving on a computer-based teletype. *Educational Studies in Mathematics, 2,* 1-15.

Tamburino, J. L. (1982). *The effects of knowledge-based instruction on the abilities of primary and grade children in arithmetic word problem solving.* Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh.

Thissen, D. (1986). MULTILOG *user's guide, Version 5.* Mooresville, IN: Scientific Software.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item-response models. *Psychometrika, 51,* 567-577.

Thorndike, E. L. (1922). *The psychology of arithmetic.* New York: The Macmillan Company.

Traub, R. E. (1983). A priority consideration in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of Item Response Theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

Traub, R. E, & Lam, Y. R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology, 36,* 19-48.

Urry, V. W. (1977). *OGIVIA: Item-parameter estimation program with normal ogive and logistic three-parameter model options.* Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center.

Urry, V. W. (1978). *ANCILLES: Item- parameter estimation program with normal ogive and logistic three-parameter model options.* Washington, DC: U.S. Civil Service Commission, Personnel Research and Development center.

Weaver, J. F. (1971). Some factors associated with pupils' performance levels on simple open addition and subtraction sentences. *Arithmetic Teacher, 18*, 513-519.

Vergnaud, G. (1982). A classification of cognitive tasks and operations of thought involved in addition and subtraction problems. In T. P. Carpenter, J. M. Moser, & T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 39-59). Hillsdale, NJ: Erlbaum.

Verschaffel, L. (1994). Using retelling data to study elementary school children's representations and solutions of compare problems. *Journal for Research in Mathematics Education*, *25*, 141-165.

Verschaffel, L., De Corte, E., & Pauwels, A. (1992). Solving compare problems: An eye movement test of Lewis and Mayer's consistency hypothesis. *Journal of Educational Psychology, 84*, 85-94.

Wheeler, L. R. (1939). A comparative study of the difficulty of the 100 addition combinations. *Journal of Genetic Psychology, 54*, 295-312.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.

Whitely, S. E. (1981). Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement, 18*, 67-86.

Whitely, S. E., & Nieh, K. (1981). *Program LINLOG.* Unpublished manuscript, University of Kansas, Lawrence, Kansas.

Whitely, S. E., & Schneider, L. M. (1980). *Process outcome models for verbal aptitude.* Technical Report NIE-80-1 for National Institute of Education. Department of Psychology, University of Kansas.

Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement, 5,* 383-397.

Willis, G. B., & Fuson, K. C. (1988). Teaching children to use schematic drawings to solve addition and subtraction word problems. *Journal of Educational Psychology, 80,* 192-201.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item-response theory.* Vancouver, BC: Educational Research Institute of British Columbia.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST *user's guide.* Princeton, NJ: Educational Testing Service.

Wolters, M. A. D. (1983). The part-whole schema and arithmetical problems. *Educational Studies in Mathematics, 14,* 127-138.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: Mesa Press.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 50,* 275-291.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432-442.